UCLA
UCLA Electronic Theses and Dissertations

Title
Essays on College Major Choice: Determinants and Centralized Mechanisms

Permalink
https://escholarship.org/uc/item/36w8x72c

Author
Ekbatani, Sepehr

Publication Date
2020

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA

Los Angeles

Essays on College Major Choice: Determinants and Centralized Mechanisms

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Economics

by

Sepehr Ekbatani

2020

© Copyright by
Sepehr Ekbatani
2020

ABSTRACT OF THE DISSERTATION

Essays on College Major Choice:
Determinants and Centralized Mechanisms

by

Sepehr Ekbatani
Doctor of Philosophy in Economics
University of California, Los Angeles, 2020
Professor Till Von Wachter, Co-Chair
Professor Rodrigo Pinto, Co-Chair

This dissertation contains three essays in applied microeconomics. The first chapter evaluates the welfare costs induced by limiting the number of choices in deferred acceptance mechanisms. I show that when the number of choices is capped, some students have to be strategic and that increasing the size of the submittable list can result in better matches, and therefore lead to welfare improvement. I use Iranian college entrance dataset to estimate a novel discrete choice model for centralized university systems, in which I relax the independence of unobserved preference shocks assumption. I validate the model with out-of-sample data from a quasi-experimental policy change, in which the list cap was increased by 50 percent. In my counterfactual analysis, I calculate that a list cap of 10 choices instead of 100 would incur a 14.2 percent welfare loss. This is equivalent to a 453 km increase in the home-university distance, which is 2.6 times the average distance traveled by Iranian students. I also show that a more restrictive list cap does not affect students at the top and bottom of the ranking, but hurts students with average scores and benefits students in the lower quartile. In the second chapter, I use the aforementioned dataset to find determinants of major choice. I estimate a rank ordered logit model of major choice and show that labor market variables, specifically earnings and unemployment play a significant role in choice of majors by students. The model shows that students prefer majors with higher expected income and expected employment rate. This study also suggests that many students care

more about the school they are applying to, rather than the major. Several explanations is possible, for example prestige of some schools might be one reason. Credit constraints that families face or the cultural barriers might also play a role for those students who prefer to stay in their hometown even at the price of studying a major they are not very interested in. Finally, in the third chapter I use neural networks to predict the number of quarters that it takes a student with certain characteristics to graduate from UCLA. I also define a survival model, in such those who did graduate before sixth year were survivors and those who couldn't were the failures.

The dissertation of Sepehr Ekbatani is approved.

Manisha Shah

Adriana Lleras-Muney

Rodrigo Pinto, Committee Co-Chair

Till Von Wachter, Committee Co-Chair

University of California, Los Angeles

2020

To Maman, Baba and Farbod.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am very grateful for the support, guidance, and encouragement of Till von Wachter, Rodrigo Pinto, Adriana Lleras-Muney, Edward Kung, and Natalie Bau. I am also thankful for the time of other faculty, especially Martin Hackmann, Manisha Shah, Moshe Buchinsky, Amir Kermani, Mohammad Akbarpour, Maurizio Mazzocco, John Asker, Simon Board, and Jay Lu for their feedback. I am grateful to my classmates, not only for comments on my research, but also for insightful academic discussions in all areas of Economics, and for making my PhD a very happy and rich experience. I am especially indebted with Carolina Arteaga, Alexandre Fon, Manos Hatzikonstantinou, Alex Graupner, Rustin Partow, Saber Ahmadi, Ehsan Azarmsa and seminar participants at UCLA, CCPR, Sharif University of Technology, and the HAND Foundation for insightful discussions. Mohammad Vesal's help was invaluable in providing access to the data. Finally, I thank my mother, father, brother, and friends, for their support throughout the entire doctoral program.

VITA

Education

2013 B.S. (Electrical Engineering), Sharif University of Technology, Tehran.

2015 M.S. (Economics), Sharif University of Technology, Tehran.

2017 M.A. (Economics), University of California, Los Angeles.

Academic Experience

2014 Research Assistant, Economics Department, Sharif University of Technology, Tehran.

2016-2020 Teaching Assistant, Economics Department, University of California, Los Angeles.

CHAPTER 1

The Cost of Strategic Play in Centralized School Choice Mechanisms

This paper evaluates the welfare costs induced by limiting the number of choices in deferred acceptance mechanisms. I show that when the number of choices is capped, some students have to be strategic and that increasing the size of the submittable list can result in better matches, and therefore lead to welfare improvement. I use Iranian college entrance dataset to estimate a novel discrete choice model for centralized university systems, in which I relax the independence of unobserved preference shocks assumption. I validate the model with out-of-sample data from a quasi-experimental policy change, in which the list cap was increased by 50 percent. In my counterfactual analysis, I calculate that a list cap of 10 choices instead of 100 would incur a 14.2 percent welfare loss. This is equivalent to a 453 km increase in the home-university distance, which is 2.6 times the average distance traveled by Iranian students. I also show that a more restrictive list cap does not affect students at the top and bottom of the ranking, but hurts students with average scores and benefits students in the lower quartile.

1.1 Introduction

Many cities and countries around the world use centralized systems to assign students to schools, and most have implemented the mechanism presented by Gale and Shapley [1962], called deferred acceptance (**DA**). There has been a rise in the universality of DA in the past two decades, and the number of students who are assigned to schools and universities under such a centralized mechanism has been growing considerably. Cities such as New York, Paris, and Madrid and countries like Norway, Chile, Turkey, Tunisia, and Iran, among others, use a similar system to assign students to secondary and postsecondary educational institutions.

The popularity of deferred acceptance is mainly due to its favorable theoretical properties. While it is not ex ante Pareto-efficient, it is well known that DA generates a *stable* matching

which is Pareto superior to all other stable matching mechanisms (Gale and Shapley [1962]). The most desirable property DA possesses is *strategy-proofness* (Abdulkadiroğlu and Sönmez [2003]), which means that players do not gain from lying about their preferences— i.e., *strong truth-telling* is the dominant strategy Nash equilibrium (Dubins and Freedman [1981]; Roth [1985]). However, all of these properties hold as long as players are not constrained in their choice making or, more concisely, the game is not a *quota game*. When agents are constrained and are only allowed to reveal a subset of their preferred choices—like most real-life implementations—the mechanism is no longer strategy-proof and it is unclear whether matching under such mechanism remains stable (Haeringer and Klijn [2009]; Fack et al. [2019]).

Despite the wide use of DA and the welfare implications of the outcomes for millions of students worldwide, only a small recent literature in economics evaluates different aspects of this mechanism in practice. The goal of this paper is to empirically study the welfare costs induced by the constraint on the list size in settings in which DA is implemented and the constraint's distributional effects across students.

To achieve this goal and to incorporate realistic assumptions on students' choice behavior, I propose a two-dimensional choice model in which each choice bundle consists of a major and a university, in which each element is valued separately by the agent. The model allows me to distinguish the demand for majors from the demand for schools, which is crucial for analyzing the choice behavior of students in college choice settings. The main contribution of this model is to relax the independence of the preference shocks assumption imposed by the usual rank-ordered logit models. Relaxing this assumption is important in those contexts in which choices are close substitutes, such as college choice settings. As an example, a student's unobservable taste for major A in school X is highly correlated with his taste shock for major B in school X or major A in school Y. The independence of preference shocks fails in other situations as well, such as settings in which students simultaneously choose high school and track or medical school and city, etc. I relax this assumption by introducing two taste shock terms that allow student's preference shocks to be correlated in two dimensions: major preferences and university preferences.

I also partially relax the truthfulness assumption. In principle, one can fully relax this assumption using the presented revealed preferences approach to derive countable moment inequalities and estimate the parameters based on the method proposed by Andrews and Shi [2013]. However, this method is computationally infeasible because of the size of my dataset and the number of covariates in the model. So instead, I partially relax the truth-telling assumption by presenting some moment equalities and show that the proposed model provides accurate predictions of the choice-making behavior of students both in and out of sample.

I use data on Iranian university applicants who are assigned through a student-proposing DA in which all universities give a fixed priority to an individual based on the student's performance on the nationwide exam.[1] I observe ordered choices submitted by more than 71,000 students who applied to enroll in Iranian postsecondary education in 2012. Students were allowed to submit an ordered preference list with up to 100 choices out of around 8,000 available options, in which each choice was a bundle of a major *and* a school (henceforth *a program*). The result is a dataset that includes more than 4 million observations at individual-program level. Additionally, I use data from a quasi-experimental policy change in 2013, in which the list cap was increased by 50 percent, to evaluate the out-of-sample performance of my model.

Using these data, I estimate a novel model of major and university choice and use the counterfactual analyses to evaluate the welfare costs of the imposed constraint in deferred acceptance mechanisms. The counterfactual results are based on exposing agents to different list caps and analyzing their choice of optimal portfolio by implementing the algorithm proposed by Chade and Smith [2006]. The main source of the welfare loss in my counterfactual analyses, is that when the list cap is more restrictive, a student submits a less diversified list which potentially generates a worse outcome for him.

A key advantage of my setting is that in addition to estimating the model, I can use a

[1]More specifically, the system is categorized as a serial dictatorship in which matching is conducted using only one side's submitted preferences. An unconstrained serial dictatorship is *obviously strategy-proof*, according to Li [2017].

policy change to obtain experimental reduced-form estimates. Reduced-form results, based on the quasi-experimental policy change in 2013, show that the list cap change from 100 to 150 generated welfare improvement, which can be interpreted as a 56 km decrease in the home-university distance traveled by the students, 31 percent of the average distance traveled by Iranian students. My counterfactual results from the model are consistent with this observation. Additionally, I find that if students were facing a list cap of 10 instead of 100, 14.2 percent of total welfare would be lost, equivalent to a 453 km increase in the distance traveled by an average student. A list cap of 15 would decrease welfare by 10 percent, which is equivalent to a 319 km increase in the distance traveled by students.

Another important result of my paper is that a more restrictive cap has a distributional effect across students and will produce winners and losers. Students whose score is slightly above average lose the most, but students in the lower quartile of the ranking benefit from a smaller cap. Overall, a smaller cap hurts total welfare as well as the fairness of the mechanism. The fact that students omit a desirable program in the presence of restrictive cap causes the system to assign that program to students with lower ranks. If the cap were not restrictive, students would apply to all desirable programs and would have no regret (known as justified envy in the literature) after the assignment is done.[2]

The paper is closely related to some recent research on school and college choice mechanisms and restrictions in DA settings. For instance, Abdulkadiroğlu et al. [2017] show that a coordinated single-offer system dominates the uncoordinated offers in NYC's high school assignment system and those students who remained unassigned under the uncoordinated system gain the most from this algorithm modification. Luflade [2018] discusses the welfare improvements caused by the sequential implementation of DA in Tunisia. She finds that updating the information set of students on the remaining vacancies leads to a considerable welfare increase, while disadvantaged students are the ones who benefit the most. In my paper, I find the welfare improvements caused by relaxing the list size constraint in DA settings.

[2]This argument is similar to the improved stability of the mechanism, but note that a mechanism can be stable but not fair to students who have done better on the exam.

Artemov et al. [2017] show that Australian students submit ordered lists that are different from their true preferences. They assert that in the case of constrained DA, skipping *out of reach* and *not good enough* options can be part of the equilibrium, so truth-telling is not the dominant strategy. Fack et al. [2019] and Agarwal and Somaini [2018] use Monte Carlo simulations to show that in constrained DA settings, estimates under a truth-telling assumption turn out to be biased.

Abdulkadiroglu et al. [2006] show that moving away from a priority matching mechanism (the Boston mechanism) to a strategy-proof one will remove the strategic burden from parents and give equal chances to both well-informed families and those who are poorly informed. This is an important policy issue, since the amount of information that families have is highly correlated with the amount of resources available to the students. Thus, any policy that removes the strategic incentives and favors unsophisticated players can lead to greater fairness and more equal access. Kapor et al. [2018] also show that moving from a strategic mechanism to a strategy-proof deferred acceptance mechanism benefits students applying to public schools in New Haven, Connecticut. In my paper, I empirically estimate the welfare consequences of strategic play in settings in which DA is constrained.

To the best of my knowledge, the only paper that estimates the list size limitations of centralized mechanisms is Ajayi and Sidibe [2015]. They use applications of high school students in Ghana to estimate a model of high school choice and evaluate the welfare costs by changing the maximum number of listings from a baseline of 6. In Ghana, however, students submit their preferences before taking the exam and before knowing their rankings, which is a deviation from the regular deferred acceptance mechanism and renders students' belief formation process very complicated. In contrast, the only source of uncertainty in my setting is the lack of information on other students' preferences which more closely represents an assessment of the classic DA. The student knows her score in the centralized exam and her priority index in the rankings, and also has access to previous years' matching outcomes. Also, the novel two-dimensional model developed in this paper provides more accurate predictions of choice-making patterns than models of major choice in the literature and is also able to capture strategic behavior in constrained school choice mechanisms.

The paper is organized as follows. In Section 2.2, I describe the data used in the study and also the institutional background of the higher education system in Iran. I develop the model and discuss the identification strategy in Section 1.3. The estimation of the model is presented in Section 1.4. Finally, Section 1.5 discusses the welfare analysis, and Section 2.4 concludes.

1.2 Data and Institutional Background

This section provides a general description of secondary and postsecondary education in Iran. Figure 2.1 provides a timeline of high school events. Details on the public and unique administrative data used in this study are also discussed.

At the end of the first year of high school, students in Iran have to choose between three broad tracks– Mathematics and Physics, Experimental Sciences, and Humanities and Literature– which will determine the set of courses they will take in the following 3 years of high school. Mathematics and Physics students will have some exclusive courses, such as Geometry, Calculus, etc. Exclusive courses for Experimental Sciences include Biology and Geology, among others, and for Humanities they include Philosophy, Advanced Literature, and others.

Students participate in nationwide diploma exams at the end of their third year of high school, which are held at the same time across the country for everyone pursuing a diploma from one of the aforementioned broad tracks. The scores on these exams will be a small portion of their final score when applying to universities.

The fourth year of high school is not compulsory for those who don't want to pursue higher levels of education. These students will receive their diplomas by passing the diploma exams and can enter the job market. Those who plan to go to university have to sign up for the last year of high school, called the *pre-university* year. Students have regular school classes for the first 6 months of the school year and have the other three to prepare for the university entrance exam, which is usually held in late June of each year.

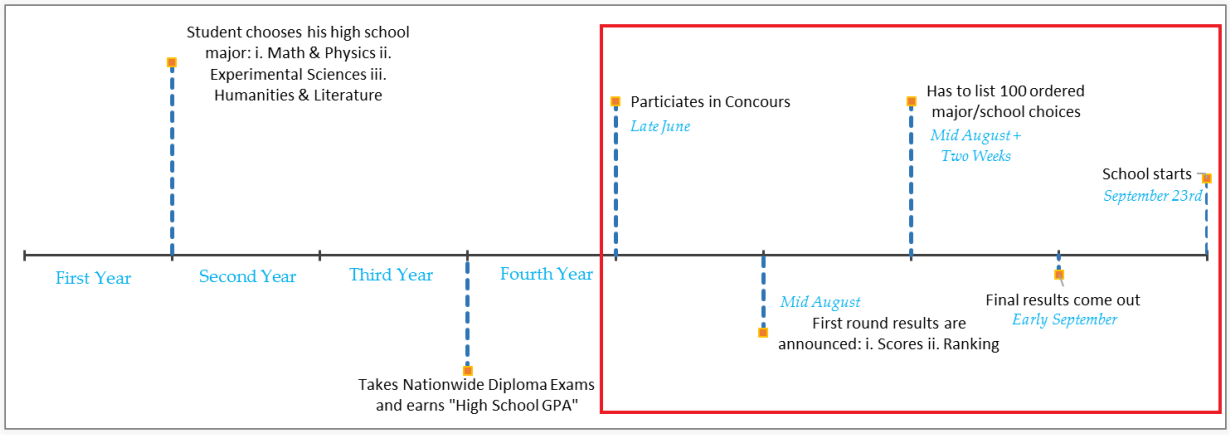


Figure 1.1: High School and Pre-University Timeline of Events

Concours is a 4-hour multiple choice exam which, for each broad track, consists of the courses students have taken throughout their 4 years of high school. Every year, around 900,000 students who have completed 4 years of high school participate in Concours. To enter a university, students must participate in Mathematics and Physics, Experimental Sciences, or Humanities Concours based on their major in high school. Participation in the Arts and/or Foreign Languages Concours is optional, and all students can take them.

After taking the Concours, students have to wait for about 45 days for the first round of results to come out, which includes their score on the exam and their ranking among all students. The final score is a weighted average of relative grades, meaning that a student will receive a higher score if the average grade of other students is very low, compared with the case in which everybody does well. Students will be ranked based on their final score to determine their priority index at the time of assignment to universities.

Along with the announcement of scores and rankings, the National Organization of Educational Testing (**NOET**) publishes a handbook containing information on programs' codes and also the number of vacancies for each program. Students have 2 weeks to fill out a form with 100 (before 2012 and 150 after) choices of major/university (each has a specific code; I call this major/university combination a *program*) and submit it to the NOET. Students can access information that provides them with previous years' results and the placement of students with ranking similar to theirs. Thus, students have some sense of the programs

they are likely to be admitted to and programs they have no chance of entering. Although 100 choices is relatively high compared with most other settings in which DA algorithm is used, it is still small compared to the total number of programs students could choose from. According to Figure 1.2, this limit seems binding for the 25 percent of students who list all 100 rows in the registration form; most of whom have done poorly on the exam.[3]

In 2012, math and physics students were allowed to choose from 8,602 different programs. Programs are very detailed in the sense that, for example, software engineering and hardware engineering (both known as Computer Engineering) have different codes and a student must apply to each separately. In total, the number of majors students had as their options added up to 241 different majors and sub-majors. The available universities numbered 854, with a minimum of three and maximum of 46 presented majors. Some of these institutions are branches of a head institution and are located in different cities, with each offering a different set of programs with different codes.

After this, it is NOET's turn to assign students to programs. The system starts with the first person in the ranking and assigns her to her desired choice. This process continues until all seats for a program are filled. From then on, students who have chosen that option as their first choice will be rejected and the system will go to their second choice on the list. The process ends when either all of the seats in all programs are filled or the last student in the ranking is assigned. There is no administrative assignment for unassigned students, and those who are left without an admission must take the exam the next year and go through the whole process again. This high cost of being left out ensures that truth-telling will not be the best strategy for students with low scores. Also, the system bans an assigned student who does not enroll from retaking the exam for 1 year, in order to ensure that students do not randomly include programs they do not like and occupy a spot that potentially could be assigned to another student.

The assignment is a one-sided Gale Shapley algorithm (serial dictatorship), in which

[3]The size of the list for the case of Chile is eight (Hastings et al. [2015a]); Norway 15 (Kirkeboen et al. [2016]); Tunisia 10 (Luflade [2018]); Ireland 10 (Chen [2012]); Turkey 24 (Saygin [2016]), etc.

universities' preferences for students are not considered.[4] Students are encouraged to list their choices in order of their preference, and since the mechanism has been in use for many years, its features are well known to the majority of students. Table 1.1 shows the summary statistics for students who took the Math and Physics Concour on June 28, 2012.

Panel A describes students' characteristics such as age, gender, city of residence, and whether they are retaking the entrance exam. Of all students who took the Concour in 2012, 60 percent were female; The share for the Math and Physics Concour was 41 percent. Students from the five largest cities (Tehran, Isfahan, Mashhad, Tabriz, and Shiraz) make up 35 percent of the sample, while students from mid-size cities are 43 percent of the sample and the rest come from small cities and villages. Of all students, 28 percent had taken the exam more than once prior to 2012.

Panel B summarizes the pattern of choices that are submitted by students. Students have, on average, filled 63.6 choices out of 100. About 26 percent of their choices were in the same city of residence, while 21 percent of their choices were for programs in Tehran. Each student has, on average, listed 17.1 different majors and 20 different universities. Panel C describes three attributes of students' first choices and Panel D shows the description of the assigned choice. Ten percent of students are unassigned, and will have to take the exam the next year in order to enter university. The rest of the students are on average admitted to their 31st choice; 24 percent are admitted to one of their first 10 options, and 3 percent are admitted to one of their last 10 options.

Table 1.2 presents some statistics for students' behavior on their submitted lists. Column (1) shows that 3.71 percent of students submitted lists with fewer than 10 choices, and 31.19 percent of students submitted lists with more than 90 choices. Column (2) shows that more than half of the students are assigned to one of their top 30 choices, and around 10 percent of students are assigned to one of the last 30 choices they submitted. As column (3) shows, choices at the top of students' lists are closer to their city of residence compared with lower-ranked choices, while column (4) shows that top choices were more selective in the previous

[4]Description of the DA is presented in Section 1.9.

year. The last column shows the share of majors at Sharif University as the most prestigious school. Note that this school has a capacity for only 885 students, which is equal to 0.34 percent of students.

As I will discuss in more detail later on, the data show that students have a preference over universities and a preference over majors and, based on these, they choose a set of programs, rank them, and submit their list. Figure 1.3 shows the patterns for the first and second choices submitted by students. Figure 1.3a shows the correlation between first and second submitted majors, and, as stated, 49.74 percent of students submit programs that share a common major, while Figure 1.3b shows that the first and second choices for 49.9 percent of students are at the same university.

Along the same line, Table 1.3 documents the fact that in the majority of students' lists, there exists one major (and/or a university) that many programs are chosen because of its desirability. Column (1) shows the share of students who applied to one major in many universities: 70.6 percent applied to one major at more than 10 universities. These students show a strong preference for that specific major, and their list conveys information on their ranking of universities. Column (2) of the same table shows that 71.9 percent of students apply to one university for more than 7 majors. Similarly, the ranking of these majors by these students helps us identify the major specific parameters of the model.

1.2.1 Policy Change in 2013

Addition to the 2012 data, I obtained access to data on students who took the Concours in 2013. A nice feature of this dataset is that it provides me with out-of-sample variations that can be used to validate my model. Students were allowed to submit lists with up to 150 choices after 2013, which was 50 more slots than 2012. The number of choices students submitted in 2013 are shown in Figure 1.4. Around 8 percent of students submitted all 150 options, while around 27 percent submitted more than 100 choices. This is additional information on the list size of 100 being binding for students in 2012. Interestingly, the list cap of 100 is 4 to 12 times the list caps in other countries with similar settings.

Table 1.4 shows students' behavior facing different caps in 2012 and 2013. Comparing students' behavior in 2012 and 2013 shows how students in 2013 were able to better diversify their submitted lists. Students in 2013, on average, submitted longer lists and listed more popular programs and more programs with lower ex ante probability of admission.

1.3 The Model

In this section, I present the two-dimensional choice model that I develop to analyze students' choice behavior. The most important property of the model is that it allows the unobservable taste shocks for programs to be correlated in two dimensions: majors and universities. With this feature, the model is able to explain the behavior of those students who apply to many majors at a specific university or one major at many universities.

There are N individuals with priority indexes, $rank_i$ $(i = 1, ..., N)$, who are allowed to list K choices in an orderly fashion from the set of J $(J > K)$ available programs, and each option is a bundle of a *major* (m) and a *school* (s): $\mathcal{J} \equiv \{(m, s) : m \in \{m_1, ..., m_M\}\ \&\ s \in \{s_1, ..., s_S\}\}$. Each element of this set has a capacity, $q(j)$ $(j \in \mathcal{J})$, of seats specific to itself, while all programs have the same preference over the pool of individuals; This is determined by the priority index.

The problem of choosing a list of programs, given the constraint on the list size, is one kind of simultaneous search problems discussed by Chade and Smith [2006] in which the cost of listing an extra choice goes to infinity after listing the K^{th} option. In these types of problems, an individual optimally chooses a portfolio from available programs that can include up to K choices, taking into account both the ex post utilities and possibility of admission.

A student receives a von Neumann-Morgenstern (**vNM**) utility, u_{ij}, by enrolling in program $j \in \mathcal{J}$, conditional on her admittance. Admission outcomes are determined by DA algorithm and after processing all the applications, but since the student is not fully informed about the preferences of other students, she will have a subjective probability in mind, p_{ij}, that represents her admission chance to program $j \in \mathcal{J}$.

The truncated DA does not allow students to list all of the available programs, so those who have more than K desirable programs for listing face a trade-off between the value they put on some of the competitive programs and their admission chances. Given the fact that a majority of students like to attend popular programs, those who are not at the top of the priority ranking have to play strategically. Otherwise, by only listing their most desirable programs, they face the risk of complete rejection by the system.

The underlying assumption is that individuals are rational agents who maximize an expected utility function for listing their preferences. The expected payoff of the list relies both on the ex post utility and the probability of acceptance of the choices. Consequently, most desirable programs might not be on the list because of their low contribution to the expected payoff, due to the low probability of admission. The following equation shows the expected utility that student i derives from list $L_i = (l_i^1, ..., l_i^k, ..., l_i^{K_i})$, where l_i^k is student i's k^{th} choice:

$$EU_i(L_i) = \sum_{k=1}^{K_i} \left[\left(\prod_{r=1}^{k-1} (1 - p_{ir}) \right) p_{ik} u_{ik} \right] + \prod_{k=1}^{K_i} (1 - p_{ik}) u_{i0} - c|L_i| \tag{1.1}$$

In this equation, K_i is the total number of listings by student i ($1 \leq K_i \leq M$), p_{ik} is the subjective probability of admission, and u_{ik} is the utility the student receives by studying at his k^{th} listed choice. The nature of the DA algorithm implies that the student will be assigned to his k^{th} choice only if he is rejected by the $k-1$ choices listed before that. This explains the second term on the right-hand side of the equation, which defines the utility the student derives from his outside option (u_{i0}) in case of getting rejected by all of the listed choices. Finally, the cognitive cost of an extra listing is denoted by c.

Each individual has a vector of vNM utilities of being admitted to programs, $u_i = (u_{i1}, ..., u_{iJ})$, a vector of subjective admission probabilities, $p_i = (p_{i1}, ..., p_{iJ})$, and a reservation utility, (u_{i0}). Using these values, a student will form the portfolio that generates the highest expected utility for her. As Proposition 4.2 in Haeringer and Klijn [2009] suggests, a student cannot do any better than to list her choices in the order of ex post utilities. This intuition follows from the assignment mechanism being deferred acceptance. A student will

get assigned to her second choice only if she is rejected by the first one, so rationality would imply that the choice with higher utility should be listed first. An important point is that the chosen programs will be sorted by utility, but information about the options that are left out stems from the revealed preferences assumption of the model. As mentioned before, the left-out options might be superior to the chosen programs, but are not listed because of the low probability of admission.

In my model, the cost of an extra listing is small as long as the number of choices has not reached the list cap. This cost can be interpreted as any nonmonetary cost, such as the cognitive cost students might incur while they are listing their choices. This marginal cost is calibrated to improve the accuracy of the model when predicting the share of students who submit a full list. Overall, a small marginal cost does not affect the counterfactual analysis, since it would be the same regardless of the list cap size.

Solving the model requires finding the utility and subjective probability vectors for each student and then the portfolio that maximizes their expected utility. As I will discuss in the next subsection, utilities can be found independently of the probability values and only by making assumptions about students' choice behavior. Assumptions on how to treat the revealed preferences and the unobservable taste shocks are the key determinants of the estimation outcomes.

1.3.1 Recovering Students' Preferences

In this subsection, using the evidence from data, I will describe the shortcomings of the usual indirect utility approach in my setting and, more broadly, in the college choice setting. Then I propose an alternative model that is based on a weaker and more plausible assumption on the unobservable taste shocks that can fit the data accurately.

A typical indirect utility function for choice models looks like the following:

$$u_{ij} = V_{ij} + \epsilon_{ij} = V(Z_{ij}, \beta) + \epsilon_{ij}. \tag{1.2}$$

This equation identifies the deterministic part of the utility student i receives from being assigned to program j, with a parametric assumption on function V and matrix of observables Z_{ij}. Randomness is introduced by adding an idiosyncratic taste shock ϵ_{ij}, which is assumed to be i.i.d with a type-I extreme value distribution.

The i.i.d assumption over i and j is common in estimation of different types of choice models. In the context of college and school choice, the i.i.d assumption over i implies that there are no peer effects and over j implies that a student's unobservable taste for one choice is independent of his taste for the other choices. I assert that this *Independence of Irrelevant Alternatives* (**IIA**) would be problematic in a college choice setting and other settings in which each choice is a choice of more than one object.

In my setting, independence of preference shocks implies that a student's unobservable taste for *UCLA Mathematics* is independent of his taste for *UCLA Statistics* and *UC Berkeley Mathematics*, which does not seem realistic, particularly with typically sparse educational administrative data. A student who has a strong preference for a major can apply to that major at many different universities. Or someone who has a strong unobservable preference for one university (like someone whose sibling is enrolled at that university) can apply to different majors in that school. This makes the choices of one student highly correlated rather than independent. As discussed in Section 2.2, looking at Figure 1.3 and Table 1.3 further suggests that choices are highly correlated.

This leads to introduction of the two-dimensional choice model. The main goal of this model is relaxing the i.i.d assumption of unobservable taste shocks in settings in which choices are correlated along several dimensions. In college choice settings, each choice of a student is in fact two choices over the pool of majors and over all universities. A student cares about the university where she will study and the major she will study separately. In my model, student i receives the following utility if she is assigned to major m at school s:

$$u_{ims} = V(Z_{ims}, X_{ms}, \beta) + \nu_{im} + \xi_{is}, \tag{1.3}$$

where V is assumed to be linear in covariates such as major and university fixed effects,

distance, distance squared, program location, and its popularity. The main property of this model is the distinction between the student's unobservable taste shock for major m, ν_{im}, and her unobservable taste shock for school s, ξ_{is}. This will allow the model to account for the correlation between taste for programs that share a common major or a common university. In the case of the aforementioned example, the student's tastes for *UCLA Mathematics* and *UCLA Statistics* are correlated through ξ_{is}, and the term ν_{im} connects her tastes for *UCLA Mathematics* and *UC Berkeley Mathematics*.

The main identification assumption is that there is no individual-major-school specific taste shock. In other words, any unobservable taste shock is either major specific or university specific and, conditional on them, preferences are explained by observable terms. This assumption seems to be supported by the data, since only 0.75 percent of submitted choices are singletons—i.e., both program's major and program's university are unique in student's list. Further, ν_{im} and ξ_{is} are assumed to have type-I extreme value distribution, but the identification strategy, described in Section 1.4, allows me to avoid putting any assumptions on the joint distributions of these taste shocks.

As previously asserted, given the choice of programs, there will be one ordered list that produces the highest expected utility and that is created by sorting the programs by their ex post utilities. The student will only take his admission probability into account when deciding on which K options to choose out of the possible J programs. Put differently, deciding which programs to list is accomplished using data on utility and probability combined, but ordering the chosen programs is only conducted by the order of utilities.

1.3.2 Revealed Preferences

In this subsection, I explain the idea behind my identification strategy to partially relax the truth-telling assumption on revealed preferences and describe the best model that can fully relax it. Section 1.10 provides a general discussion of different approaches to using revealed preferences to estimate model parameters and different implications of each assumption. I will discuss that the best model cannot be estimated because of computational difficul-

ties. Instead of full relaxation, I describe my strategy to partially relax the truth-telling assumption and explain the underlying intuition.

Fack et al. [2019] and Agarwal and Somaini [2018] argue that the truth-telling assumption is unrealistic under a constrained DA, and show that estimates under this assumption turn out to be biased. A less limiting assumption about students' decision-making behavior is *undominated strategy*, which only assumes that students do not play dominated strategies. This assumption implies that the submitted list should be sorted by the order of preference, and no information is obtained from left-out choices. In other words, in equilibrium student will submit a *partial preference order* of the options he finds both desirable and feasible given his priority. This approach by students results in a not unique, but an undominated strategy Nash equilibrium in which the student submits an ordered list of those programs he thinks he has a chance of getting into. Under the undominated strategies assumption, j is revealed preferred to j' if the former is ranked higher on the list. The implication of such an assumption about observing such ordering can be written as

$$\begin{aligned} Pr\left(j \succ_i j'\right) &= Pr(u_{ij} > u_{ij'} \text{ and } j, j' \in L_i) \\ &\leq Pr(u_{ij} > u_{ij'}). \end{aligned} \tag{1.4}$$

My proposed identification strategy is to use revealed preferences on majors and universities separately. Comparing two majors that are listed in the same university allows me to focus only on the unobservable taste shocks for majors. A similar argument about comparing two universities with the same major holds for the unobservable taste shocks for universities. The following definitions shed more light on the identification strategy.

Definition 1 *: Major m_1 is revealed preferred to m_2 at school s if program (m_1, s) is listed higher in ranking compared to (m_2, s):*

$$\begin{aligned} Pr(m_1 \succ_{i|s} m_2) &= Pr(u_{im_1s} > u_{im_2s} \cap (m_1, s), (m_2, s) \in L_i) \\ &\leq Pr(u_{im_1s} > u_{im_2s}) \end{aligned} \tag{1.5}$$

Equation 1.5 yields a lower bound for $Pr(u_{im_1s} > u_{im_2s})$, while the following shows a higher bound for the same term:

$$1 - Pr(m_2 \succ_{i|s} m_1) \geq Pr(u_{im_1s} > u_{im_2s}). \tag{1.6}$$

The probability on the right-hand side of both equalities only consists of major-variant terms; the school-specific terms cancel out. Similarly, the next definition compares choices with a common major that will omit the major-specific terms.

Definition 2 *: School s_1 is revealed preferred to s_2 for major m if program (m, s_1) is listed higher in ranking compared to (m, s_2):*

$$\begin{aligned} Pr(s_1 \succ_{i|m} s_2) &= Pr(u_{ims_1} > u_{ims_2} \cap (m, s_1), (m, s_2) \in L_i) \\ &\leq Pr(u_{ims_1} > u_{ims_2}). \end{aligned} \tag{1.7}$$

And for the higher bound:

$$1 - Pr(s_2 \succ_{i|m} s_1) \geq Pr(u_{ims_1} > u_{ims_2}). \tag{1.8}$$

Using these two definitions, the model is identified by the distributional assumption on each error term, ν_{im} and ξ_{is}, separately. The right-hand-side probabilities will take a logistic form, assuming that the error terms have a Gumbel distribution. Estimation based on this assumption is more complicated than the alternatives because of the introduction of inequalities. In Section 1.4, I will describe how one can use moment inequality methods to estimate the parameters of the choice model, assuming undominated strategies.

Strong truth-telling (**STT**) is known as a strong assumption, while its weaker form, *weak truth-telling* (**WTT**), receives a lot of attention in the school and college choice literature.[5] Undominated strategies, as the weakest assumption on students' choice behavior, seems

[5]For instance, Drewes and Michael [2006]; Hastings et al. [2009]; Hällsten [2010]; Kirkebøen [2012]; Budish and Cantillon [2012]; De Haan et al. [2015]; and Luflade [2018].

the most appealing one in this context. Unfortunately, estimation based on undominated strategies is not computationally straightforward, and often yields wide and informative bounds on coefficients (Fack et al. [2019]).

The computational intensity of estimating the model based on moment inequalities forces me to make further assumptions and estimate the model based on moment equalities. If inequalities in Equation 1.5 - Equation 1.8 are replaced by equalities, they imply that students are truthful about the majors they list in a given school, or about the universities that they list for studying a given major.

Although this sounds like general truthfulness, it is a partial relaxation of the regular truth-telling assumption. If a student has listed major A at school X but has not listed major B at school Y, truth-telling implies that program (A,X) is preferred to (B,Y). However, the assumption I am making does not put any restriction on the preference over these two programs that do not share a common major or a common university. Still, assuming equalities is a stronger assumption compared with undominated strategies, because it assumes that all of the majors that are not listed by the student at a given school are less preferred to majors that are listed at that school. Also, all of the schools that are not listed for a given major are inferior to those schools that are listed for that major.

1.3.3 Recovering Subjective Admission Probabilities

The second component of the expected utility equation in Equation 1.1 is the subjective admission probabilities. How students form their expectations about their chance of admission matters for the choice of options that will ultimately be listed.

I assume that the student has access to historical public data on admission thresholds and how variant those thresholds have been over the years. For that purpose, I use a representative sample of students who has listed a program over the course of 7 years prior to the year of this study to run a *limited dependent variable* model and to estimate the probability of admission, given the ranking of each student. Student i computes his probability of getting admitted to program j by estimating the share of students who had his ranking on the exam

and were accepted to the program. The following equation,

$$P(\text{Accepted to } j|\text{Rank} = r_i) = F_j(r_i), \tag{1.9}$$

can be estimated using logit. The subjective probability of admission given student i's ranking will follow:

$$p_{ij} = \hat{F}_j(r_i). \tag{1.10}$$

The data and fitted values are shown for a selected number of programs in Figure 1.7. This figure shows four different programs, from the most popular program to some not so popular ones that almost all students have some chance of getting admitted to. I used data on both students who were assigned and those who showed interest on their list and had a higher rank than the threshold, but were assigned to a higher choice on their list.

Another approach could be to examine the cutoffs over the years and find the fitted probabilities based on potential acceptances. My approach takes into account the number of students who have shown interest in addition to the assigned students. The logit estimates will be determined both by the pattern of thresholds and also the average number of students applying to that program.

1.4 Estimation and Results

Based on the assumption on revealed preferences and the unobservable taste shocks I explained in the previous section, in this section I will describe the identification strategy and provide the results on students' preferences. These results, along with the estimated subjective probabilities, allow me to use the model in the counterfactual analysis that will follow. I will also provide results based on truth-telling and the normal i.i.d taste shocks to highlight the fact that these assumptions yield biased estimators that cannot be reliable.

The two inequalities provided by Definition 1 can be written in moment terms as follows:

$$\begin{aligned} Pr(u_{im_1s} > u_{im_2s}|Z_{im_1s}, Z_{im_2s}, \beta) - \mathbb{E}\left[\mathbb{1}(m_1 \succ_{i|s} m_2)|Z_{im_1s}, Z_{im_2s}\right] \geq 0 \ ; \\ 1 - \mathbb{E}[\mathbb{1}(m_2 \succ_{i|s} m_1)|Z_{im_1s}, Z_{im_2s}] - Pr(u_{im_1s} > u_{im_2s}|Z_{im_1s}, Z_{im_2s}, \beta) \geq 0 \ . \end{aligned} \tag{1.11}$$

For every school, there are two moment conditions for each pair of majors that are listed by students. In total, there will be M conditional moment inequalities obtained based on these sets of comparisons. On the other hand, Definition 2 can be used in the same way to find S conditional moment inequalities based on the school comparisons. Writing Equation 1.7 and Equation 1.8 in moment terms:

$$\begin{aligned} Pr(u_{ims_1} > u_{ims_2}|Z_{ims_1}, Z_{ims_2}, \beta) - \mathbb{E}\left[\mathbb{1}(s_1 \succ_{i|m} s_2)|Z_{ims_1}, Z_{ims_2}\right] \geq 0 \ ; \\ 1 - \mathbb{E}[\mathbb{1}(s_2 \succ_{i|m} s_1)|Z_{ims_1}, Z_{ims_2}] - Pr(u_{ims_1} > u_{ims_2}|Z_{ims_1}, Z_{ims_2}, \beta) \geq 0 \ . \end{aligned} \tag{1.12}$$

These inequalities are interacted with Z_{ims} to obtain $M + S$ unconditional moment inequalities. To estimate Equation 1.3, one can follow the approach proposed by Andrews and Shi [2013] and construct the following objective function $T_{MI}(\beta)$ based on the inequalities in Equation 1.11 and Equation 1.12:

$$T_{MI}(\beta) = \sum_{j=1}^{M_1} \left[\frac{\bar{m}_j(\beta)}{\hat{\sigma}_j(\beta)}\right]_-^2 , \tag{1.13}$$

where $\bar{m}_j(\beta)$ and $\hat{\sigma}_j(\beta)$ are the mean and the standard deviation of the j^{th} moment and $[a]_- = min\{0, a\}$.

Similar to Fack et al. [2019], estimations based only on the inequalities yield uninformative bounds in my setting. To achieve point identification I replace the presented inequalities with equalities; I discussed its implications in Subsection 1.3.2. The estimation results based on these assumptions are shown in Table 1.5. For comparison, I have included the estimates obtained by commonly used rank-ordered logit method in **??**. It is clear that rank-ordered logit estimates are biased, as some other papers have shown. When the regression is run

without the university fixed effects, the coefficient for distance turns out to be positive. This means that if a program is moved 100 km farther from the student, the log odds of it being chosen goes up by 1.2 percent. This is inconsistent with what is documented in Table 1.2, that students prefer programs that are closer to their hometown. This positive coefficient flips sign when university fixed effects are added to the regression, but still shows a downward bias compared with the same setting in column (2).

Focusing on the second column of Equation 1.3 shows that students dislike distance and 2-year programs. Closer universities are more important to female students and slightly more to students from large cities. Students strongly prefer universities that are located in their city and less strongly universities that are located in their providence of residence. There is a strong preference for programs that are in Tehran, but most of it is captured when the fixed effects for famous universities (such as Sharif; Tehran etc.) are taken into account. Coefficients on major fixed effects are relative to a village male student's preference for a major in humanities. Engineering, architecture and civil engineering are the most popular majors, followed by computer ccience. The change in the number of observations is because some observations appear in major equalities and also in school equalities. So it is possible that each program in the original setting appears up to two times in the estimation by the proposed model.

Using the parameters in Table 1.5, I find the flow utility that each individual receives from all programs. Some programs generate positive and some negative utility for students. Based on the vector of all flow utilities and the vector of all subjective probabilities that I found in Subsection 1.3.3, in the next section I describe how I find the portfolio with the highest expected utility from the pool of all programs, given different list size caps.

1.4.1 Model Fit

Finding the optimal portfolio is not a computationally easy task, because of the large number of programs an individual can choose from. For instance, for an individual who chooses 100 programs out of almost 8,000 options, the number of potential portfolios is on the order

of 10^{232}, which is infeasible to solve. To be able to solve the model, I use the algorithm presented by Chade and Smith [2006], the *marginal improvement algorithm* (**MIA**). The authors prove that to reach the optimal portfolio, one must, at every step, pick the next best choice that contributes the most to the expected utility of the existing list.

MIA runs in the following steps:

1. Start with $L_i = \emptyset$; Discard all alternatives with flow utility less than the outside option (u_{i0}).

2. The program with the highest expected utility $(p_{ij}u_{ij})$ is chosen first: $L_i = \{j_1\}$.

k. Select the best complement to the current list L_i:

$$\max_{j_k} \quad EU(L_i')$$

$$L_i' = \text{arranged elements of } (L_i \cup \{j_k\}) \text{ in decreasing order of utility.}$$

The algorithm starts with an empty portfolio, then finds the program that gives the highest expected utility— i.e., the program with the highest interaction of utility and subjective admission probability. After that, it finds the program that improves the expected utility the most, with the consideration that programs should be ordered by utility. In other words, the student has the option to choose a program to add to the top (Extension), bottom (Insurance), or middle (Diversification) of his list. He might choose a top school that he loves but for which he does not have much of a chance. He might add a safe option to the bottom of his list, which he does not prefer over all other options, but adding it will reduce the risk of getting rejected by the system. Or he might just add to the diversity of his portfolio by adding another option to the middle of his list; still, this option will be added after the ones that dominate it. The improvement will depend on where the program is added because of the interactions of the $(1 - p_{ir})$ terms in the expected utility function described in Equation 1.1.

Chade and Smith [2006] prove that the marginal improvement algorithm will yield the optimal portfolio for the student. The implication of this method is that programs are not

chosen in the order of utility but in the order of the improvement they make to the expected utility of the list. If the list size the student can submit is increased, the student will add a desirable choice with a nonzero chance of admission to his list. This can potentially change the outcome of his assignment in equilibrium, given the lists submitted by other students. This is the main source of the welfare improvement observed, which will be discussed shortly.

I assume that the outside option for students provides them with zero utility (which can be easily relaxed), and I calibrate the marginal cost of an additional listing (c) to fit the fact that 25 percent of students submit a full list. I expose students to a list cap of 100 and compare the predicted behavior with the observed data. Figure 1.8 suggests that with $c = 10^{-8}$, model prediction fits the number of students who submit a full list. The model predicts that 25 percent of students submit a full list, which is equal to the observed 25 percent.

The model does well when it comes to predicting students' behavior throughout their submitted list. Figure 1.9 shows that students list programs they have around 10 percent chance of getting into as their first choice. They list less popular programs with higher ex ante probability of admission as they move down their list. This observation is captured pretty well by the model, since the numbers are very close to the real data. The only difference is the last couple of listings, where model students act more conservatively and the probability of acceptance shoots up. This can be explained by the one-time shot that students in the model get versus the option of taking the exam the next year for students in the real world.

The policy change in 2013 provides me with extraordinary information to validate my model out of sample. The parameters of my model are estimated based on data from 2012, when the list cap was 100. If I expose the students in my model to a list cap of 150, they will submit a different composition of programs. Comparing the predicted behavior with the actual data in 2013 can provide information on how good the model can predict students' behavior when they are facing caps other than 100. Similar to Figure 1.9, the prediction of the model and the 2013 data are shown in Figure 1.10. The figure shows that the model has a very decent performance, even out of sample, in terms of predicting students' choice

behavior.

1.5 Welfare Analysis

In this section, the welfare gain from changing the list cap in DA is evaluated. Welfare is defined as the following uniformly weighted utilities of the students, where (m^*, s^*) is the program the student is assigned to:

$$W = \sum_{i=1}^{N} u_i(m^*, s^*). \tag{1.14}$$

Note that the utility here is the ex post utility the student receives after her assignment is completed. This welfare measure can be defined with unequal weights to also include the fairness of the assignment. In what follows, I evaluate the welfare effects of the policy change from 2012 to 2013, and then describe the counterfactuals I run using my model and discuss their welfare implications.

1.5.1 Quasi-experimental Policy Change

In 2013, students were allowed to submit 50 more choices compared with students who took the exam in 2012. Figure 1.4 depicts the distribution of the number of choices submitted by students in 2013. It shows that around 8 percent submitted the full 150 choices, and around 27 percent submitted a list with more than 100 choices. This number is close to the number of students who were facing a list cap of 100 in 2012 and submitted a full list.

Using the parameters in Table 1.5, which are based on the data from 2012, the flow utility of students' assigned choices is calculated for both 2012 and 2013. Figure 1.11 shows that the flow utility of students has improved as a result of the policy change. This result shows that welfare is increased by an equivalent of 56 km after the policy change. This number is close to the median of distance traveled by students, which is 63 km.

This can be seen as evidence on how changing the list cap can affect total welfare. The

cohort in 2013, who were allowed to make 150 choices, were able to submit a more diversified list, which benefited them. This can be seen by comparing Figure 1.9 and Figure 1.10. While the patterns through the lists are close to each other, students in 2013, on average, list programs in which they have a lower chance of admission compared with students who submitted their applications in 2012. Interestingly, the rate of rejection doesn't change from 2012 to 2013, so the change in welfare is only due to better matches for students who were facing a binding list of 100 in 2012.

1.5.2 Counterfactual Results

Using the utility and probability values, I expose students to different list limitations. When students are playing a quota game they do not list their most preferred choices; instead, their first pick will be the option that gives them the highest expected utility. The student will continue by choosing the option that gives her the most marginal expected utility, given her previous list up to the point that she runs out of choices or the cost of an extra listing dominates the benefit. Another way of putting the same concept is to ask her to remove options from her submitted list to reach a list with a more binding cap. She will not simply omit her last options; instead, she gives up some of the diversification of her list. This removes her chance of getting into a program for which she had a minimal admission chance. This arises because the ex ante probabilities of admission are not taken from a complete information set on the submitted lists of other students. In this situation, removing some options might change the final outcome and ultimately hurt the student.

In response to different list sizes, students will submit different lists; thus, a different equilibrium is expected. Submitted lists in each round will be fed to the DA algorithm to find the final allocation of students and also the students who remain unassigned. Each match generates some utility for assigned students, and unassigned students will receive their reservation utility. I use Equation 1.14 to calculate the total welfare under each counterfactual and show the results in Figure 1.12. I show the total welfare calculated for each list size and normalize the result with respect to the list size of 100.

On the other hand, Figure 1.12 shows that given my model, an increase of 50 choices from 100 to 150 should generate 0.89 percent improvement in total welfare. According to the results in Table 1.5, this welfare improvement is equivalent to a 27 km decrease in the home-university distance for an average student. Comparing this result with the reduced-form results shows that, as discussed before, the model produces a lower bound for the welfare gain, since all agents in my model are well informed and sophisticated.

The main result of this paper is that total welfare would be 10 percent lower if students were facing a list cap of 15, and 14.2 percent lower if the cap were only 10. Ten percent lower welfare is equivalent to a 319 km increase in distance traveled by an average student, which is double as the average distance an average student travels in Iran.[6] In the case of the list cap being equal to 10, the average student will lose utility equivalent to traveling 453 km more. This considerable welfare mainly comes from the worse matches that students receive when the cap is more binding. Figure 1.13 depicts the welfare change in terms of change in the distance traveled by an average student.

An important implication of my results is that a less binding cap will produce winners and losers. Intuitively, Figure 1.14 shows that top- and low-ranked students are not affected by the increase in the list cap from 10 to 100. Most importantly, this figure shows that students who have a moderate score benefit the most from a less binding list, which provides them with a better match, while students who are in the 20th to 40th percentiles lose some of their utility. The reason behind this is that some programs were not listed by students in the middle of their ranking because of the binding list they were facing, and initially these programs would go to the lower-ranked students. This phenomenon does not happen when the cap is 100 and the middle-ranked students will fill the programs they are qualified for and receive a higher utility from them.

This sheds light on the fact that in addition to improving welfare, a less binding cap will add to the fairness and meritocracy of the matching—higher fairness in the sense that students who have done better on the exam and come out higher in the ranking receive

[6]The average distance traveled by students is 176 km and the median is 63 km.

better matches.

1.6 Conclusion

In this paper I developed a two-dimensional discrete choice model for college major choice to evaluate the welfare costs of list truncation in deferred acceptance settings. I estimated the model using a rich dataset from Iran, in which students are allowed to make up to 100 choices out of around 8,000 options available. I moved away from the usual assumptions in the literature; specifically, I relaxed the independence assumption on individual taste shocks for programs and partially relaxed the truth-telling assumption. I showed that many students, mostly those with low ranks, play strategically and are not truthful. I also showed that students have either a stronger preference for the university or the major of study. In this situation, the independence of unobservable taste shocks is unrealistic, and the model needs to be built on different set of assumptions.

For demand estimation, I developed the true model based on moment inequalities— but, due to computational complexities, estimated the model using moment equalities. I showed that the usual rank-ordered logit model generates biased estimators that cannot be reliable. The estimated model was used in counterfactual analysis to find the welfare cost of strategic play in DA systems. To find the optimal portfolio under different list caps, I used the marginal improvement algorithm proposed by Chade and Smith [2006].

My results show that a more binding implementation of DA reduces welfare considerably. Having a list cap of 10 instead of 100 reduces welfare equivalent to a 453 km increase in the home-university distance traveled by students, which is three times the average distance students travel. I showed that increasing the list size from 100 to 150 has a positive welfare impact that is close to the impact of other modifications proposed in the literature, such as sequential implementation of DA.

I showed that changing the cap does not affect two groups of students, those at the top and those at the bottom of the ranking. However, I also showed that a less binding mechanism produces a more fair matching, which can be tied to the stability of the mechanism. In a

binding mechanism, students give up desirable choices that might be taken by a student with a lower rank.

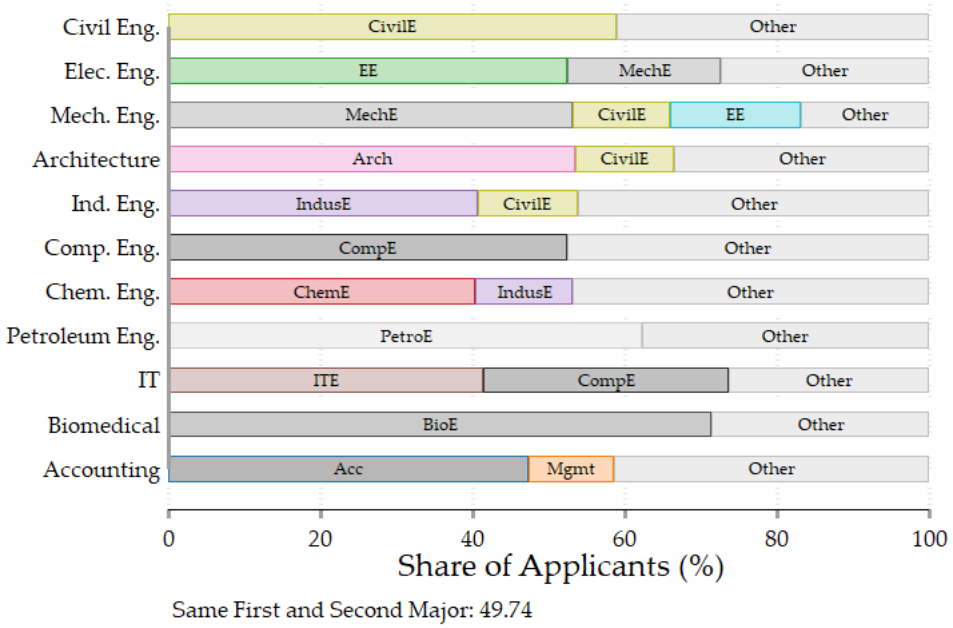
Increasing the list cap may be the most efficient and cheapest modification of systems in which truncated DA is implemented. The costs of such modification are mainly the information processing cost and the cognitive cost of listing additional choices. In the current era, the former cost seems negligible compared with the benefits of such improvement. The cognitive cost of listing more choices is also dominated by the cost of strategic play in not strategy-proof settings.

1.7 Figures

Figure 1.2: Number of Listings

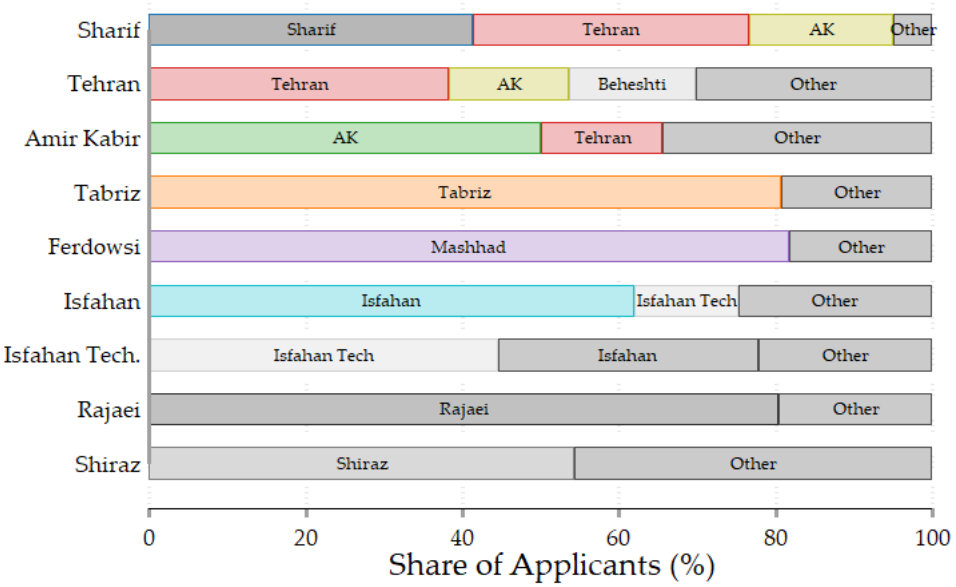
Note: Histogram of the number of submitted choices by students. At 2012, the cap on list size was 100 and students were allowed to submit a list up to size 100. The graph shows that around 25 percent of students submitted a list consisting of 100 options and that the cap had been binding for one in every four students.

Figure 1.3: Second-choice Program by First-choice Program



Same First and Second Major: 49.74

(a) Second-choice Major by First-choice Major

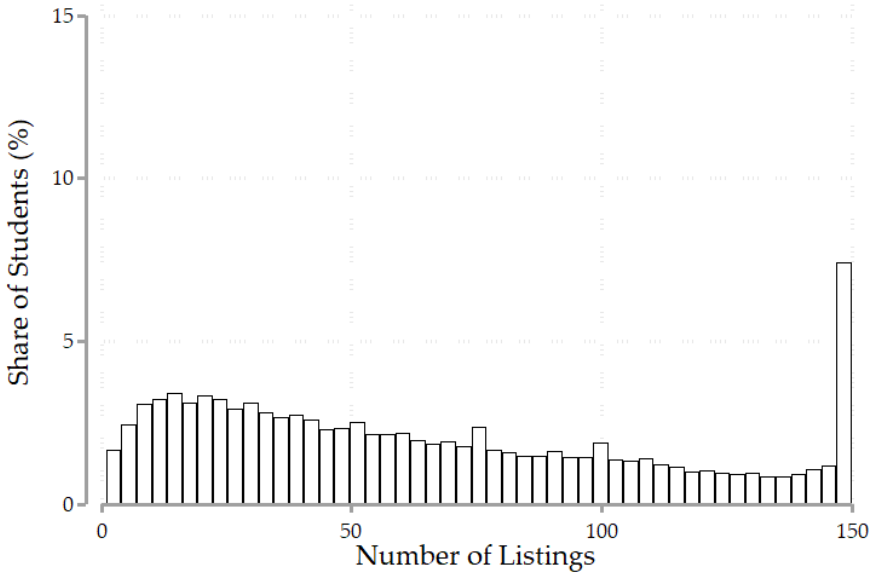


Same First and Second University: 49.9

(b) Second-choice University by First-choice University

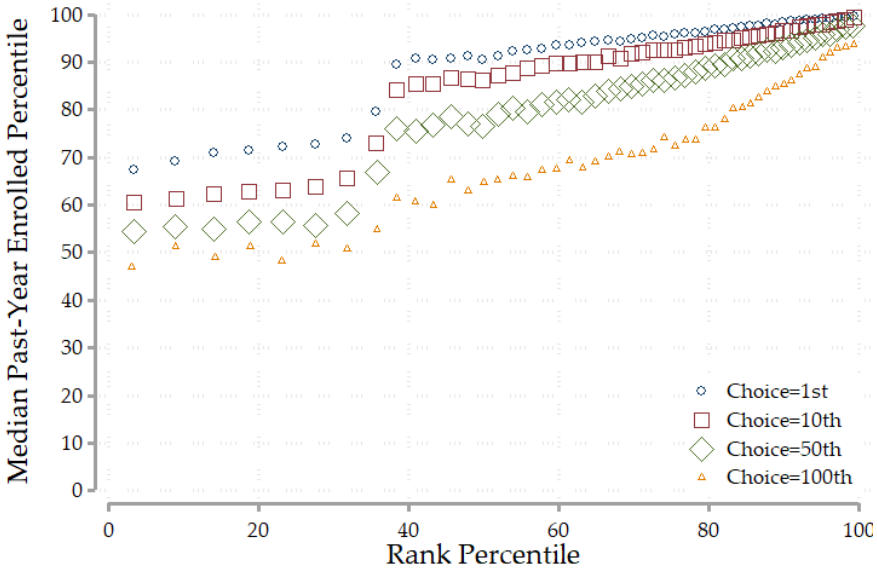
Note: (a) shows the share of choice number 2's major categorized by student's first chosen major. For 49.74 percent of students, choices 1 and 2 share a common major. (b) shows the share of choice number 2's university categorized by the first listed university. 49.9 percent of students listed the same university in their first and second choices. The other 0.36 percent of students chose their first two options, such that they do not share either a common major or a common university. (*Other* includes all major/universities whose share was less than 10 percent.)

Figure 1.4: Number of Listings Histogram After Policy Change



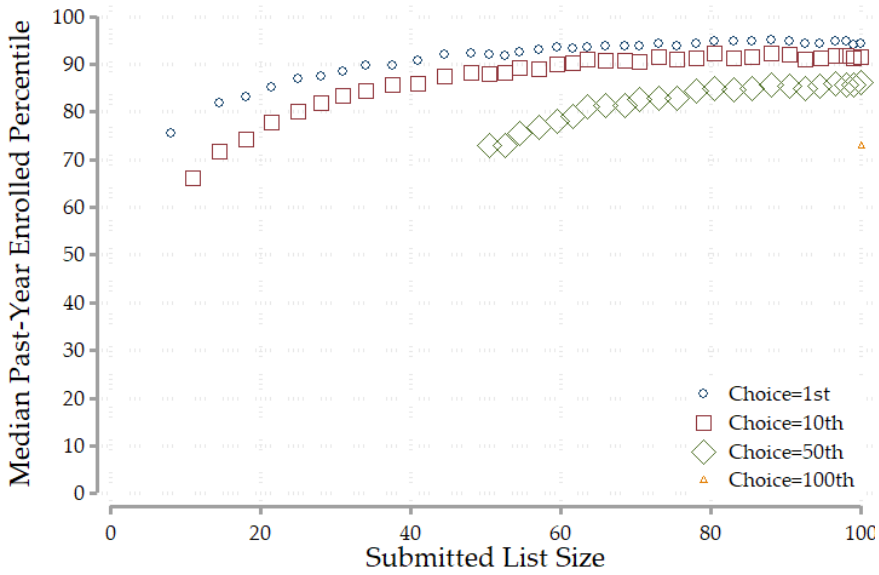
Note: This graph is similar to Figure 1.2 for 2013 and after the change in the cap from 100 to 150. Note that the scales are different, but it seems that almost the same number of people who were listing 100 in the previous setting are listing 100 or more choices. This could be evidence on the list of 100 being binding for that 25 percent of students in 2012.

Figure 1.5: Selectivity of Choices by Student Ranking



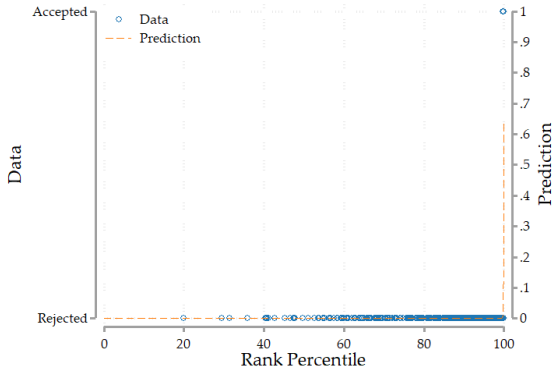
Note: This bin-scatter graph shows the choice behavior of students with different ranks in terms of popularity of the programs. Students are sorted on the x-axis from the one with the lowest score to the top student, who is in the 100^{th}percentile. On the y-axis, programs are sorted by their selectivity, proxied by the median rank of the students who enrolled in that program the year before. Students with higher ranks have listed very selective programs, even as their 100^{th}choice. On the other hand, low-ranked students have not listed very selective programs because of their close-to-zero chance of admission. This figure is provided as evidence that DA is not strategy-proof when agents face a list cap.

Figure 1.6: Selectivity of Choices by Submitted List Size

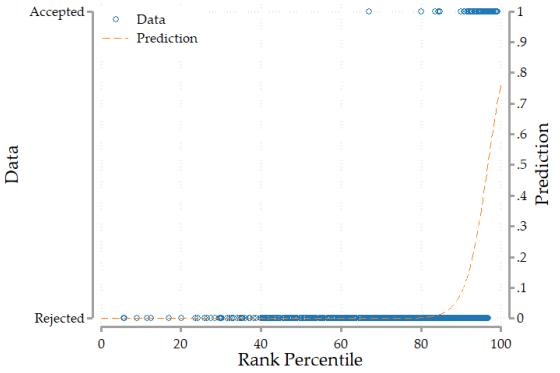


Note: This figure shows that students who submitted a list with fewer than 100 choices (a *short list*) are not necessarily truth-telling. On average, students who submitted a short-list have chosen less selective programs as their 1^{st}choice (blue circles), 10^{th}choice (red squares), etc. This graph provides evidence against the common claim in the literature that short-list students are truth-telling.

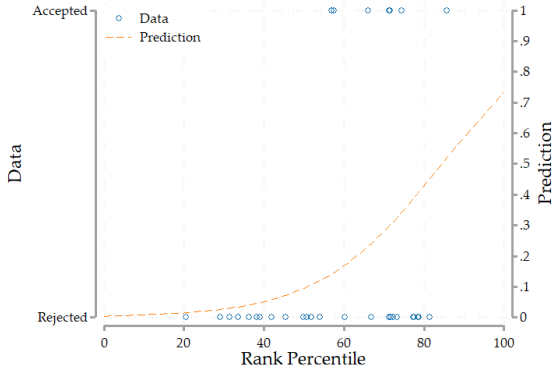
Figure 1.7: Historical Data on Acceptance by Rank Percentile and Logit Prediction



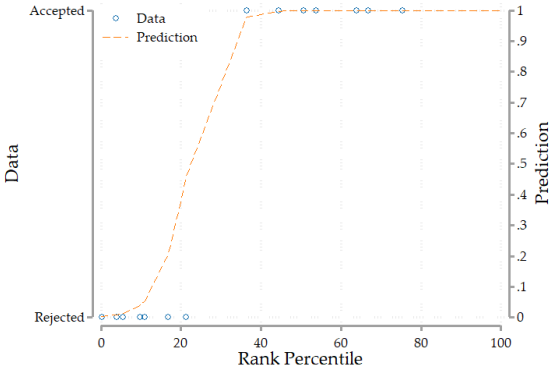
(a) Electrical Engineering, Sharif University of Technology, Tehran



(b) Industrial Engineering, Bu-Ali Sina University, Hamedan



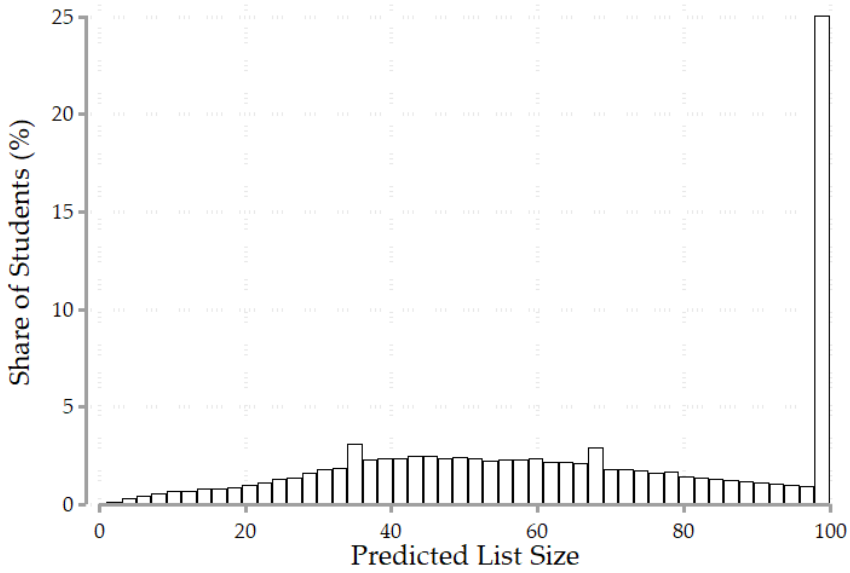
(c) Physics, Lorestan University, Khorram Abad



(d) Accounting, Payam Nour University, Bostan Abad

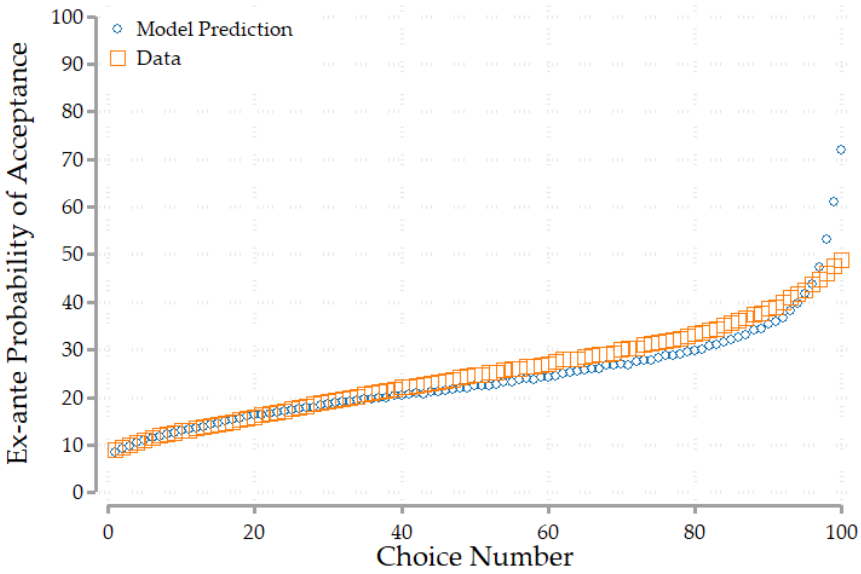
Note: Example of the estimation procedure of the subjective probabilities for four different programs. Students who listed the program over the 7 years prior to the study are sorted on the x-axis based on their ranking, and their result (Accepted or Rejected) is shown as zero or one circles. With logit estimation, the probability of getting accepted given the rank is predicted and used in expected utility Equation 1.1.

Figure 1.8: Predicted List-size Histogram



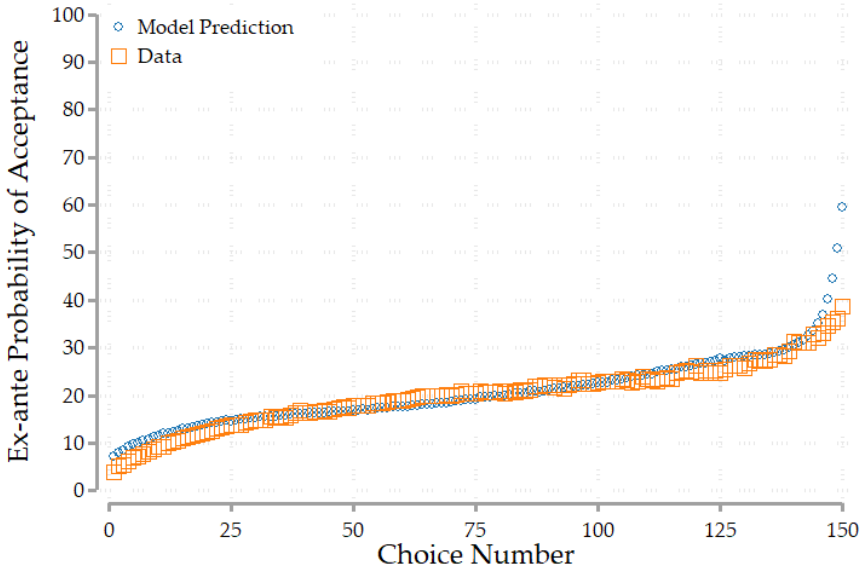
Note: Histogram of number of listings predicted by the model when the list cap is 100. Students stop listing if they have an option with probability 1 in their list, or every available option is less desirable than the outside option or the expected utility improvement of an additional choice is less than the marginal cost. With $c = 10^{-8}$, the share of students who submitted a full list matches the data. For comparison, see Figure 1.2.

Figure 1.9: Ex ante Admission Probability of Listings
Data and Prediction



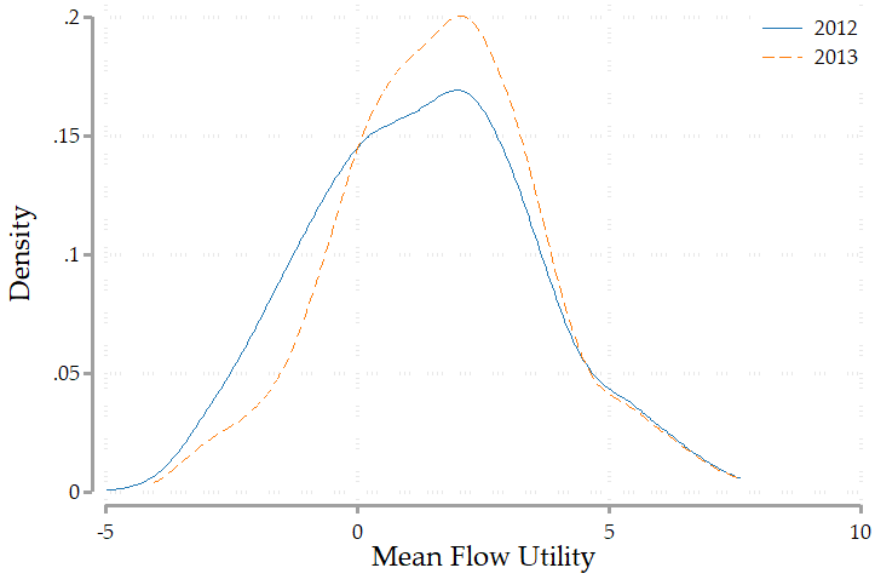
Note: Ex ante probability of admission throughout the list. The first choices on the student's list are on average those for which the student does not have a very high chance of admission. As she moves down the list, she lists programs that are more accessible, although they might be less popular. The model predicts this pattern of students' behavior pretty well. The last four choices by the students in the model are much more conservative than those in the data. The reason is the absence of a retaking option in the model compared with real life.

Figure 1.10: Out-of-sample Prediction of the Model



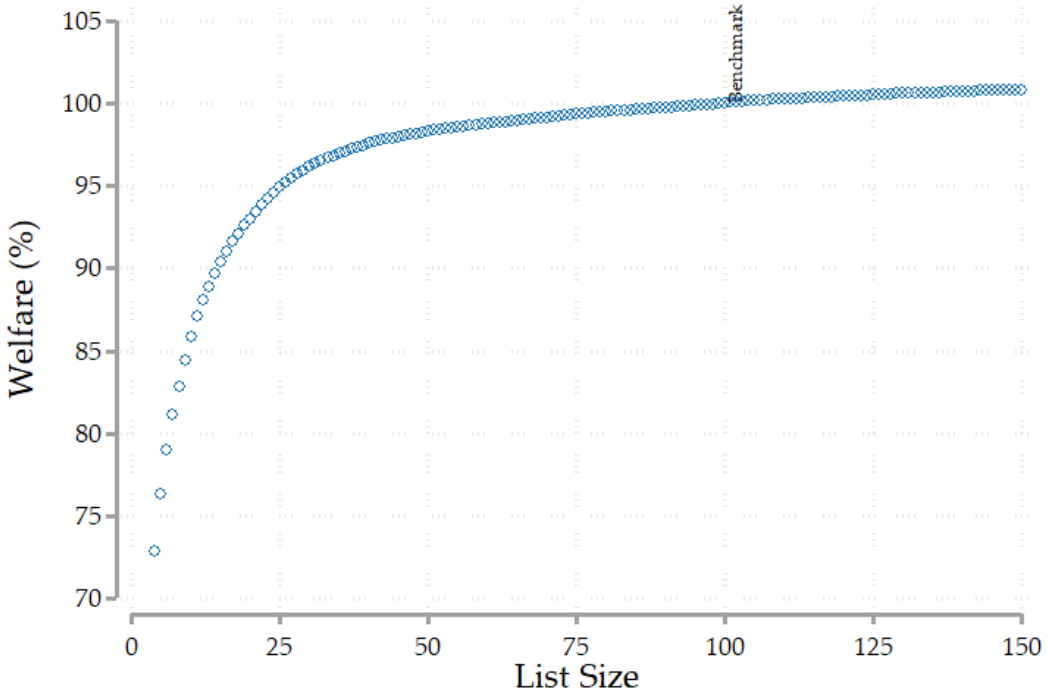
Note: Ex ante probability of admission throughout the list. The model prediction is based on data on applications in 2012 and the data is on applications in 2013. This figure shows the validity of the model out of sample. The first choices on the student's list are on average those for which the student does not have a very high chance of admission. As she moves down the list, she lists programs that are more accessible although they might be less popular. The last choices by the students in the model are much more conservative than in the data.

Figure 1.11: Demeaned Flow Utility Before and After Policy Change



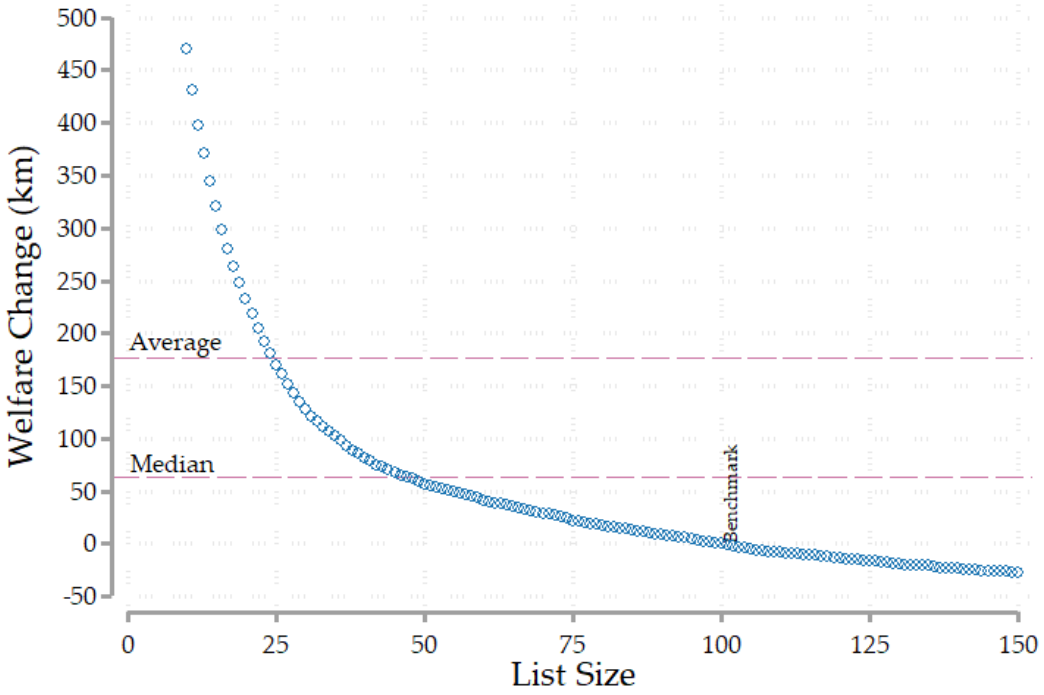
Note: This graph shows a reduced-form result for the welfare improvement of increasing the list size under DA. The solid blue line shows the density of mean flow utilities for the assignments before the policy change (list size equal to 100), and the orange dashed line shows the density after the policy change (list size equal to 150). The distribution has shifted to the right, which suggests welfare improvement.

Figure 1.12: Counterfactual Welfare Analysis. List Size of 100 is the Benchmark.



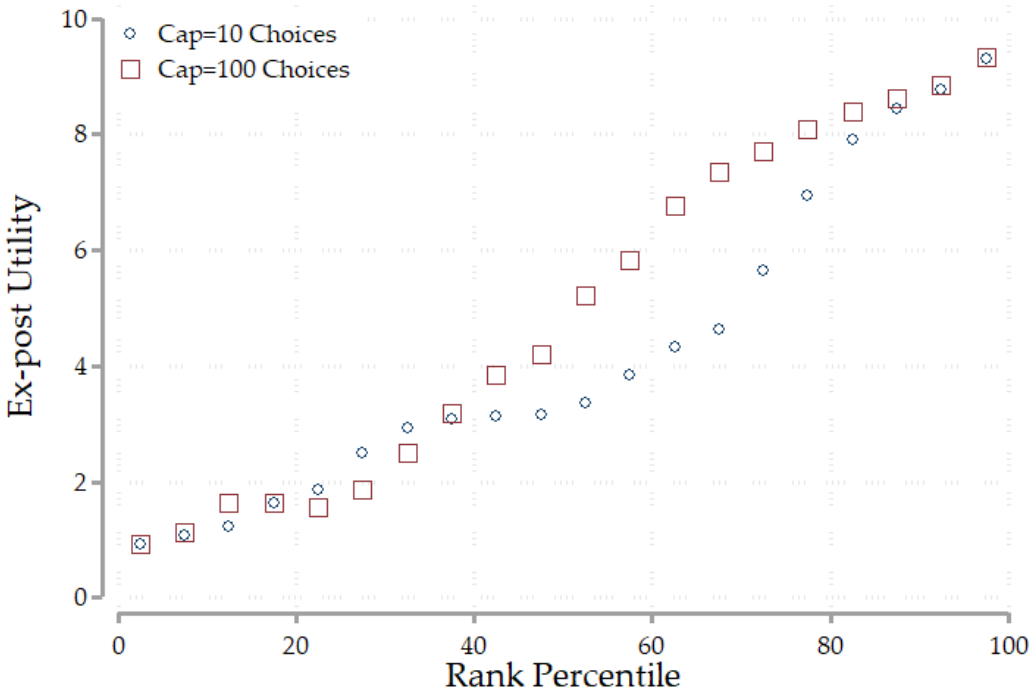
Note: Counterfactual analysis of the model. Total welfare calculated by Equation 1.14 is shown on the y-axis for different list sizes. Each student submits a different list when facing a different list cap, which will result in a different outcome for him. The sum of the utilities is the measure of welfare under different list sizes.

Figure 1.13: Counterfactual Welfare Analysis in Distance Terms. List Size of 100 is the Benchmark.



Note: Welfare change is translated to change in distance traveled by an average student. The benchmark is the list cap of 100. Average distance traveled by Iranian students is 176 km, and the median is 63 km. These statistics are shown with dashed lines on the graph.

Figure 1.14: Winners and Losers



Note: Utility derived from the assigned choice by the student's rank for cap sizes of 10 and 100. Students at the top and bottom are not affected by the change in the cap, while for students in the middle, some benefit and some lose. As the list cap increases, students are able to submit a better diversified portfolio, which will provide them with a more desirable outcome. This takes away the chance of getting into those programs from some students with lower ranks. This can be interpreted as increasing the fairness of the mechanism.

1.8 Tables

Table 1.1: Total Summary Statistics

Variable	Mean	Std. Dev.	Min	Max
Panel A. Student Characteristics				
Age	19.11	1.86	16	59
Female	0.41	0.49	0	1
Large Cities	0.35	0.47	0	1
Mid-size Cities	0.43	0.49	0	1
Small Cities and Villages	0.22	0.40	0	1
Concour Rank	85,380	61,364	8	229,910
Retaking the exam	0.28	0.45	0	1
Math score (%)	12.76	15.68	-21.2	98.2
Physics score (%)	13.48	16.20	-24.4	97.1
Chemistry score (%)	16.83	17.29	-26.6	97.2
Panel B. Choices				
Number of Listings	63.66	30.84	1	100
Choices within same city of residence	0.26	0.31	0	1
Choices at Tehran	0.21	0.27	0	1
Majors Ranked (Total=241)	17.12	9.48	1	58
Universities Ranked (Total = 854)	20.02	13.19	1	94
Panel C. First Choice				
Distance (km)	214.2	319.8	0	2480
Tehran	0.36	0.48	0	1
Same city as residence	0.35	0.47	0	1
Panel D. Outcomes				
Rejected	0.10	0.30	0	1
Row of accepted choice	30.75	25.94	1	100
Distance (km)	176.2	280.1	0	2480
Same city as residence	0.34	0.47	0	1
Row of accepted choice $\in$ [1,10]	0.24	0.43	0	1
Row of accepted choice $\in$ [11,20]	0.14	0.35	0	1
Row of accepted choice $\in$ [21,30]	0.11	0.31	0	1
Row of accepted choice $\in$ [91,100]	0.03	0.15	0	1
Number of Students	71,918			
Total Number of Observations	4,461,572			

Note: This table provides summary statistics for students who took the Iranian nationwide university entrance exam in 2012 and their submitted choices of programs to the National Organization of Educational Testing (NOET). The sample includes 71,918 students out of 260,055 students who took the Concours on June 28, 2012. Values are shares unless stated otherwise.

Table 1.2: Students' Choice-making Behavior

n	Share of students with a list of size n or shorter	Share of students who are assigned	Mean Distance (km)	Previous Year Percentile	Share of choices at Sharif
	(1)	(2)	(3)	(4)	(5)
1	0.1	5.03	214.26	91.17	11.40
2	0.27	8.04	219.55	90.17	6.13
3	0.49	10.77	229.01	89.52	4.71
4	0.75	13.31	235.02	89.02	4.20
5	1.11	15.64	239.89	88.51	3.35
6	1.54	17.81	246.12	88.21	2.90
7	2.02	19.86	251.09	87.88	2.88
8	2.54	21.79	256.14	87.60	2.57
9	3.06	23.75	259.12	87.28	2.32
10	3.71	25.57	261.81	87.17	2.34
20	11.11	40.63	287.98	85.76	1.54
30	19.72	52.40	306.87	84.86	1.13
40	28.44	61.88	322.21	83.79	0.81
50	36.98	69.45	328.13	82.78	0.73
60	45.45	75.46	331.42	81.54	0.54
70	53.48	80.36	333.64	80.05	0.43
80	61.03	84.03	326.97	78.41	0.36
90	68.81	87.09	304.70	75.39	0.41
100	100	89.65	252.14	69.97	0.73

Note: Numbers are in percentages unless stated otherwise. This table provides statistics for major choice behavior of students in 2012. Column (1) shows the share of students who have submitted a list of n 100 or shorter. Column (2) shows the mean distance between student's residence city and the university for all students who have submitted the row. Column (3) is the share of students who submitted a choice up to a row number. Column (4) is a measure of selectivity of the choice. It shows the precentile rank of the last student who was assigned to the choice in 2011. Column (5) is the share of students who submitted their choices at Sharif University of Technology as the most prestigous school for math and physics students. (Total capacity at Sharif was equal to 885 seats or 0.34% of all students.)

Table 1.3: Nestedness of Choices

		share of students who applied to *a major* in n or more universities (%) (1)	share of students who applied to *a university* for n or more majors (%) (2)
	1	100	100
	2	99.12	99.26
	3	96.96	96.86
	4	94.01	93.48
	5	90.47	87.48
	6	86.81	80.86
	7	82.95	71.89
	8	79.07	63.5
	9	74.86	54.15
n	10	70.58	46.51
	20	30.18	7.27
	30	10.96	1.14
	40	3.94	0.20
	50	1.61	0
	60	0.68	0
	70	0.31	0
	80	0.18	0
	90	0.09	0
	100	0.01	0
	Average	5.07	3.06
	Median	3	2

Note: This table shows the correlation between students' choices. Column (1) shows the share of students who revealed their preference to study a major at many universities. For instance, it shows that 66% of students have a major in their list which they have applied for at more than 10 universities. Column (2) shows the same statistic for the choices that share a common university.

Table 1.4: Choice Behavior Comparison between 2012 and 2013

		n					
	Year	1	10	30	50	100	150
Share of students who submitted the choice number n (%):							
	2012	100	94.4	79.2	62.6	24.8	
	2013	100	92.8	76.6	59.4	27.1	7.1
Share of students who stopped before the choice number n (%):							
	2012	0.10	3.71	19.72	36.98	100	
	2013	0.21	4.23	21.75	39.53	72.95	100
Mean distance from the chosen program (km):							
	2012	187.6	265.5	332.8	364.6	297.9	
	2013	199.1	255.3	306.0	341.6	394.2	347.8
Share of students who are assigned to a program in their list before choice number n (%):							
	2012	4.89	24.91	51.04	67.65	87.33	
	2013	4.96	24.9	49.76	64.38	81.47	86.54
Share of students who submitted Sharif in their list before choice number n (%):							
	2012	11.1	17.0	19.9	20.9	22	
	2013	14.6	22.41	26.3	27.6	28.8	29.2

Note: This table provides statistics on the choice behavior of students in year of 2012, and 2013. In 2012 students were allowed to submit a list with up to 100 options, whereas in 2013 they were allowed to submit up to 150 options. The first two rows show the share of students who submitted the specific choice on the column header. The second two rows show the distribution of maximum number of listings in 2012 and 2013. As shown, more than 27 percent of students submitted a list with 100 or more options in 2013. The third two rows show students' choice behavior in terms of home-university distance. The fourth two rows show the distribution of students' accepted choice for those who are accepted. The fifth two rows show the share of students who ranked Sharif university as the most prestigious school in 2012 and 2013.

Table 1.5: Utility Parameters Estimation Results

	(1)		(2)	
Distance (100km)	-0.0493***	(0.000)	-0.148***	(0.000)
× Female	-0.0154***	(0.000)	-0.0152***	(0.000)
× Mid Cities	0.00391***	(0.000)	0.00318***	(0.000)
× Large Cities	0.0233***	(0.000)	-0.00546***	(0.000)
Distance (100km) Sq.	0.000545***	(0.000)	0.00560***	(0.000)
Past-Year Median Admit	5.039***	(0.000)	2.044***	(0.000)
2-Year Program	-1.088***	(0.000)	-1.841***	(0.000)
Same City	0.217***	(0.000)	0.346***	(0.000)
Same Province	-0.105***	(0.000)	0.116***	(0.000)
Location: Tehran	0.829***	(0.000)	0.150***	(0.000)
× Female	-0.00887	(0.053)	0.130***	(0.000)
× Mid Cities	0.0544***	(0.000)	0.184***	(0.000)
× Large Cities	-0.296***	(0.000)	0.00768	(0.321)
Major Fixed Effects	X		X	
× Female	X		X	
× Mid Cities	X		X	
× Large Cities	X		X	
University Fixed Effects			X	
Observations	7,453,671		7,453,671	

p-values in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Estimation of utility parameters by proposed model. Column (1) only includes major fixed effects and the interaction terms with student characteristics. Column (2) adds the university fixed effects to the model. *Distance* is proxied by the distance between student's city of residence and the city the university is located in. Students are divided into three categories: from villages, from mid-size cities, and from large cities (Tehran, Isfahan, Shiraz, Tabriz, Mashhad). *Distance sq.* is square of distance. *Past year median admit.* is the measure of the program's popularity, since more popular programs tend to be filled with students with higher rankings. Broad major fixed effects are included in all columns according to *ISCED97*. Column (2) includes university fixed effects; some institutions that are branches of the same university are grouped together. The variation in these groups allows the estimation of variable *Location: Tehran*.

Appendix

1.9 Description of Deferred Acceptance

Round 1. All students are tentatively assigned to their first submitted choice and sorted based on their score on the exam. The lowest-ranked students who are excess to the capacity of the program are rejected and moved to the pool of unassigned students. Others are temporarily kept for the next round.

Round k. The system will look at the next choice of those students who were rejected in the previous round and update each program's tentative list. The list will be a pool of students who were on the list in Round (k-1) and those who are added in this round. If the program has more students on the list than its capacity, the lowest-ranked students will be rejected and become unassigned students.

Stop. The system will stop when there is no unassigned student or all of the choices of unassigned students are filled.

1.10 Revealed Preferences Assumptions

The difference in the approaches comes from how different papers treat the left-out options and also the assumption they put on idiosyncratic taste shocks. Following Fack et al. [2019]'s notation, I discuss different assumptions that can be put on a student's program selection behavior and then describe the model I develop to recover students' preferences.

Strong truth-telling (**STT**) is a Nash equilibrium under the original deferred acceptance mechanism. When DA is unconstrained, students do not benefit from manipulating their true preference list or dropping any option, so STT suggests that in (a not unique) equilibrium, they will list all the available choices in the order of their true preference. The not-uniqueness of STT is due to *irrelevance at the bottom* and *skipping the impossible* behavior of students.[7]

[7]For illustration see Example 1 in Fack et al. [2019].

Assuming STT in the setting of this study is not realistic, since it requires students to list all of the available options while they face a constraint on the size of the list they can submit.

Weak truth-telling (**WTT**) is a weaker version of STT and assums that the student lists the first K options of her true preference list truthfully. This implies that the student's list starts with the most desirable choice, and every option is preferred to the next one on the list. Additionally, every choice that is not listed by the student is less preferred than listed options, and not listing them is either because of listing cost or inferiority to the outside option. Most papers that deal with discrete ordered-choice problems tend to assume some version of WTT behavior by decision makers, since it is not as strong as STT and it results in tractable closed-form solutions for the maximum likelihood function. The following describes the probability of observing the student's true list:

$$Pr\left(L = [l_1, ..., l_k, ..., l_K]\right) = \ Pr(u_1 > u_2 > .... > u_K > u_j \ : j \notin L) \tag{1.15}$$

With a specific distributional assumption on the error term of the utility function, the right-hand side probability would yield a closed-form formula that can be estimated using maximum likelihood, and this is the main reason for the popularity of this assumption in the literature.[8]

Figure 1.5 shows that truth-telling might be consistent with the choice behavior of top students, but it does not hold for low-ranked students. This figure sorts all of the programs on the Y-axis in terms of the popularity among the previous year's applicants and sorts the current year's students based on their ranking on the X-axis, with 1 being the person with the highest priority index. In that graph, popularity is measured by the median percentile rank of students who enrolled in that program. This figure shows that those who are not at the top of the ranking seem to be omitting the most popular schools because of their close-to-zero chance of admission. This can be evidence against WTT, since the not-top students do not necessarily prefer their listed choices to the ones they have left out.

[8]For instance, Drewes and Michael [2006]; Hastings et al. [2009]; Hällsten [2010]; Kirkebøen [2012]; Budish and Cantillon [2012]; De Haan et al. [2015]; and Luflade [2018].

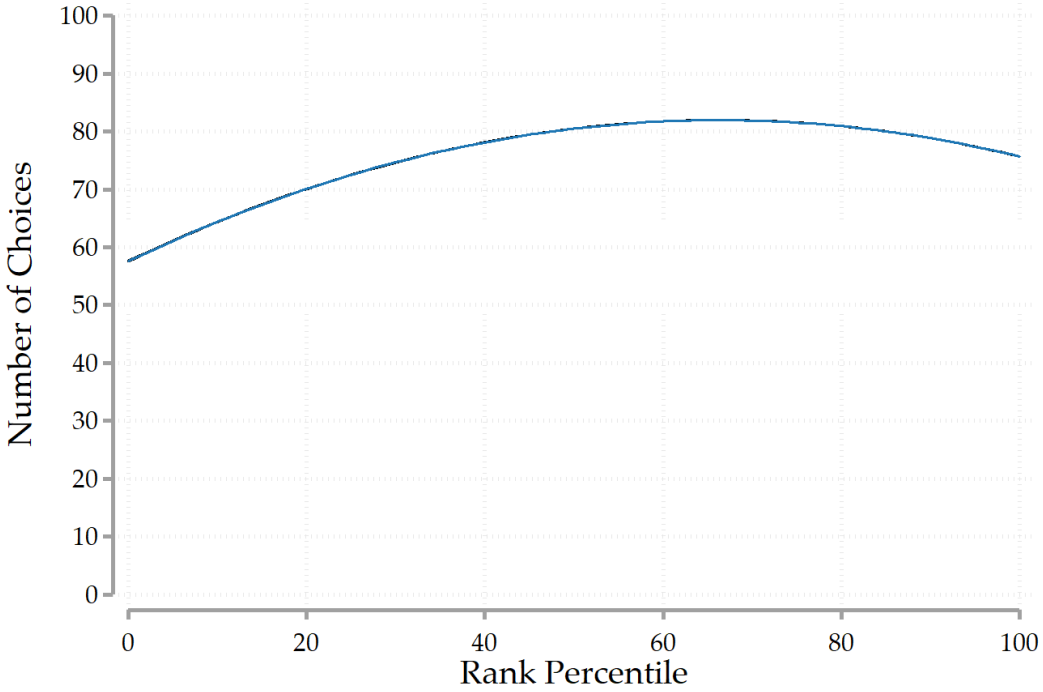
Figure 1.6 shows that assuming WTT for students who have listed fewer than 100 options is also not justified. Students who have listed fewer options seem not to be truthful, in the sense that they do not submit the most popular programs. Thus, using data from these students and generalizing it to other students might also be misleading. Based on the model, listing fewer than the cap size can be a result of (i) having zero subjective probabilities for left-out options and/or (ii) having a safe choice with admission probability equal to 1 that dominates every not-listed option with a positive subjective probability of acceptance and/or (iii) having the outside option, which dominates all the not-listed options with a nonzero subjective probability. At any rate, these students are presumably selected and also may not have representative preferences.

A less limiting assumption about students' decision-making behavior is *undominated strategy*, which only assumes that students do not play dominated strategies. This assumption implies that the submitted list should be sorted by the order of preference, and no information is obtained from left-out choices. In other words, in equilibrium a student will submit a *partial preference order* of the options he finds both desirable and feasible given his priority. This approach by students results in a not unique but an undominated strategy Nash equilibrium in which the student submits an ordered list of those programs they think they have a chance of getting into. Under the undominated strategies assumption, j is revealed to be preferred to j' if the former is ranked higher on the list. The implication of such assumption about observing such ordering can be written as

$$\begin{aligned} Pr\left(j \succ_i j'\right) &= Pr(u_{ij} > u_{ij'} \; and \; j, j' \in L_i) \\ &\leq Pr(u_{ij} > u_{ij'}). \end{aligned} \tag{1.16}$$

Estimation based on this assumption is more complicated than the alternatives because of the introduction of inequalities.

Figure 1.15: Number of Choices by Rank of Students



Note: Number of listed choices by students with different ranks.

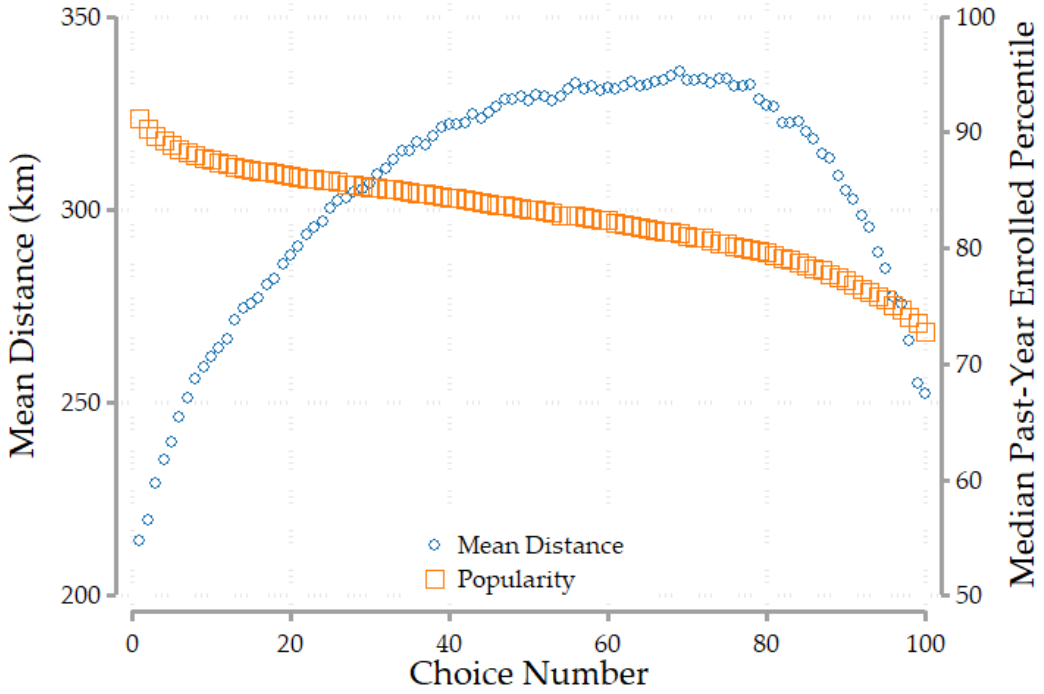
1.11 More on Choices and Data

The number of choices students with different ranks made follows an inverse U shape. Students who are top in the ranking do not need to list many choices to be assigned to a choice. On the other hand, since students who are assigned to a program are banned from taking the exam the next year, students who have done poorly on the exam and want to have the option of retaking the exam avoid submitting a full list with some random programs. The students in the middle of the ranking are the ones for whom submitting a full list matters the most. This is shown in Figure 1.15.

Students show strong preference for popular programs and programs close to their city of residence. They also choose their safe options from programs that are not popular but are located close to their hometown. Figure 1.16 shows students' choice behavior throughout their list. The trend is very consistent for the choice of popular programs, but not for distance. Closer programs are definitely more appealing to students, but only up to a point.

After students start choosing the safe options, they choose them from closer schools.

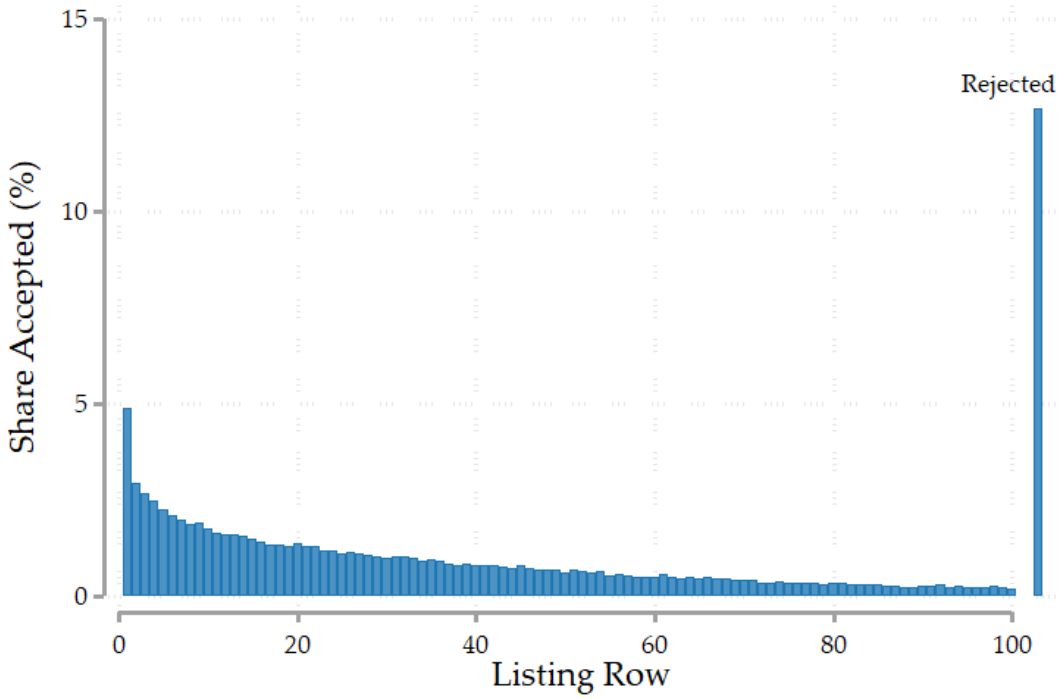
Figure 1.16: Choice Behavior, Distance, and Popularity



Note: Distance and popularity of choices throughout student's list.

As another descriptive result, I look at the share of students who were assigned by their listing number. In 2012, 10.2 percent of applicants were rejected by all of their choices and were left unassigned. Of those who were accepted to a program, students received their 30th choice on average. Figure 1.17 shows the share of accepted students for each listing number.

Figure 1.17: Share of Accepted Students by Listing Number



CHAPTER 2

The Determinants of College Major Choice: Evidence from Iranian University Entrance Exam

2.1 Introduction

There is a growing literature on the returns to education by major of study (Hastings et al. [2013],Kirkeboen et al. [2016], Altonji and Zimmerman [2017], Arcidiacono [2004]). However, we don't know much about the determinants of major choice and why students prefer one major over the others. Although several studies tried to answer this question by running surveys on college students about the characteristics of their favorite major, they lack a good data and also students are surveyed after they have chosen their major. (Arcidiacono et al. [2010], Wiswall and Zafar [2015], Baker et al. [2017])

These papers are based on questionnaires in which students are asked about their expected salary in future and comparing answers to actual labor market data to draw conclusions on lack of information and errors in expectations (Arcidiacono et al. [2010], Hastings et al. [2015b], Wiswall and Zafar [2015], Baker et al. [2017], Montmarquette et al. [2002]).

Another set of papers have found that non-pecuniary factors are the main determinants of major choice, while expected earnings play a significant but small role (Beffy et al. [2011], Wiswall and Zafar [2015], Stinebrickner and Stinebrickner [2013]). They find factors such as prestige, curriculum offerings, location, ... as the main non-pecuniary determinants of major choice. Because of data limitations, both set of papers have to look at broad major categories such as social sciences, engineering, etc. But the problem is that there is a lot of variance in both pecuniary and non-pecuniary factors among different engineering majors,

for example.

By using a large dataset on ordered major choices and preference questionnaires, this study tries to fulfill the gap in the literature. Main feature of this dataset is that it includes the list of 100 ordered major/university choices for about 100'000 Iranian students. In addition to the list of ordered choices, students are asked about their main priorities for choosing a major and their expectations about future earnings of top majors.[1] Using a revealed preference approach important factors in major choice can be determined.

The article is organized as follows. In Section 2.2, I will describe the secondary and post-secondary system of education in Iran and also the features of administrative and public data that I will use. Empirical strategy and results are discussed in Section 2.3. Conclusion and further work are discussed in Section 2.4. In Section 2.5, some facts using the public data are shown and the other possible quesions that can be addressed are discussed.

2.2 Higher Education in Iran and Data

This section will provide a general description of secondary and post-secondary education in Iran. Figure 2.1 provides an overview of sequence of high school events. Details on the public and unique administrative data that is used in this study will also be discussed.

The problem with studies on major choice that are based on surveys and questionnaires is that results are usually influenced by the fact that the student has been already admitted to the school for following a specific major and it might influence his answers to major choice preferences. So, the ideal dataset for this project will include data on not only the major that student is admitted, but also the places and majors that he wanted to get into but he got rejected. There is a valuable dataset on students who take the university entrance exam (called Concours) in Iran which has all the properties needed to find the determinants of major choice which I will discuss in short.

At the end of the first year of high school, students have to choose among three broad

[1]This administrative dataset is not public and needs to be accessed on-site.

majors, i.e. Mathematics and Physics, Experimental Sciences, Humanities and Literature. This will determine the set of courses that they will take in the following three years of high school. Mathematics and Physics students will have some exclusive courses such as Geometry, Calculus, ... Exclusive courses for Experimental Sciences include Biology, Geology, ... and for Humanities include Philosophy, Advanced Literature, ... Students in each broad major will be examined on general courses and on their own exclusive courses later on. This allows private high schools to specialize in one of the broad majors while public high schools are required by law to offer all three majors for their students which results in a lower quality.

Students will have to participate in nationwide diploma exams at the end of third year of high school. Exams are held at the same time all around the country for everyone pursuing a diploma from one of the aforementioned broad majors. The scores on these exams will be a part of the final score for entering the university later on.

Fourth year of high school is not compulsory for those who don't want to pursue higher levels of education. They will receive their diploma by passing the diploma exams and can join the labor force. Those who plan to go to university have to sign up for the last year of high school, called pre-university year. Students will have regular school classes for the first six months of the educational year and have the other three for preparing for the important exam which is usually held in late June of each year.

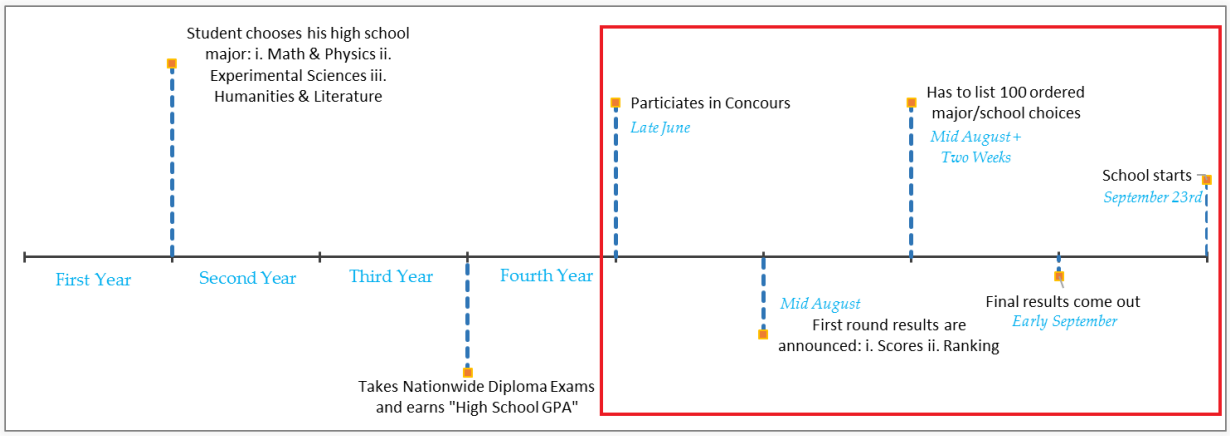


Figure 2.1: High School and Pre-University Timeline of Events

Concours is a 4-hour multiple choice exam in which for each broad major consists of

the courses that students have taken throughout the four years of high school. Every year around 900'000 students (around 60% girls), who have completed four years of high school education participate in Concours. For entering university, students have to participate in Mathematics and Physics, Experimental Sciences or Humanities Concours based on their major in high school. Participation in Arts and/or Foreign Languages Concours is optional and all students can participate in them.

Pre-university is the most important grade for all students in Iran. Private school tuition costs for this grade exceeds an average worker's yearly income. In Tehran only, there are about 1800 test preparation institutions providing mainly afternoon classes, tutoring and mock exams. Revenue of these institutions is estimated to be around $2B (Almost 0.5% of Iran GDP).

As another evidence on importance of higher education, Table 2.1 compares the number of students per 100'000 in Iran with some similar countries in terms of GDP per capita along with United States and England.

	1975	1985	1995	2000	2005	2010
Iran	428	398	1557	2191	3103	5217
Turkey	690	870	2043	2548	3147	4147
Egypt	1053	1711	-	3779	3309	3337
Brazil	1039	-	-	1629	2503	3158
United States	4730	5158	5442	4735	5905	6673
England	1267	1807	3167	3478	3844	3969

Table 2.1: Number of College Students per 100'000
Source: UNESCO Education and Literacy Statistics

After taking the Concours, students have to wait for about 45 days for the first round of results to come out. Students will know their score on the exam which is a weighted average of their relative scores on different courses on the exam. Relative score means that student will receive a higher score if the average among other students is very low. Based on these scores students will be ranked and the ranks will also be reported to the student. This ranking is the determinant of who gets to choose first his desired major of study.

After the announcement of scores and rankings, students will have two weeks to fill out a

form with 100 choices of major/university (each major/university has a specific code, I call it course from now on) and submit it to the National Organization of Educational Testing (NOET). These choices are ordered and students would not benefit from listing any major other than their true preference. Most of students ask for consulting which will provide them access to the previous years results and the information on placement of students with similar rankings as theirs. So essentially they have a probability distribution of getting accepted to each major when they are making their choices. Since there is 100 rows to be filled they will have the option to choose their desired course even if they had no chance of getting in last year with the same score.

After this period, it is NOET's turn to assign students to the majors. System will start with the first person in ranking and will assign him to his desired choice. This process continues until the seats for a course is filled. From then on the students who has chosen that option as their first choice will be rejected and system will look at their second choice on the list. This process ends when either all the seats of all majors are filled or the last student in the ranking is reached.

The large number of choices that students can list make sure that the process is a one-sided Gale Shapely algorithm which is strategyproof. It proves that people will reveal their true preferences when they are making their choices. This is actually one the main differences between Iran's system with countries which are using similar centralized university entrance exams like Norway and Chile. In these two countries number of choices that student can make is limited to 15 which makes it hard to ensure the strategyproofness. The other difference is that universities, both public and private don't decide on whom to accept and everything is done only through the NOET system.

The data on the number of seats for each major/university, rank and score of students who got accepted and their city of residence is public. The focus of this study would be on the administrative data which the list of 100 choices that students submit to the NOET in addition to many other individual characteristics. The data is classified and can only be accessed on site. Total summary statistics are shown in Table 2.2.

I have also used data from Household Income and Expenditure Survey (HIES) and Labor Force Survey (LFS) both conducted by Statistical Center of Iran to extract occupational earning and unemployment for years 2009-2015. Occupations and university majors are linked through International Standard Classification of Education 1997 (ISCED97) and Selected Characteristics of Occupations 1988 (SCO88). Table 2.7 shows all the ISCED major categories used in this study and examples of SCO occupations linked to them. Earnings and Unemployments are averaged with weights to form average salary and average unemployment of different categories. This is done for two reasons: i) Limited number of observations for some subfields in the data could have hurt the results of the paper. ii) Large number of categories make the estimation of the model very complicated and even impossible, so using these standards reduce the dimensions of the model and make it easier to solve.

Year	2009	2010	2011	2012	2013	2014	2015	2016
Number of Students	80'652	175'880	205'983	178'820	96'400	102'020	75'013	104'634
Number of Exam IDs	80'652	212'079	245'975	210'424	117'486	125'882	95'984	133'281
Average Number of Exams per Student	1	1.21	1.19	1.18	1.22	1.23	1.28	1.27
Average Number of Listings per Student	67	68	64	60	66	68	73	68
Total Number of Observations	5'401'227	14'367'653	15'663'283	12'576'816	7'719'203	8'604'346	7'010'853	9'048'905

Table 2.2: Total Summary Statistics

2.3 Empirical Strategy and Results

In order to find the determinants of major choice, I start with a simple OLS approach in which an index of the popularity of the major is produced and then regressed on associated labor market outcomes. I calculate the average ranking of students studying different majors and then regress this average ranking on labor market variables (Columns (1) and (2) of Table 2.3). This index might be affected by the capacity of majors in the way that if a major is only thought in the best school of the country then it would go first in the ranking and the average score of other majors would be affected by the students' scores of other schools and other cities. I define another index of popularity by counting number of top 10'000 students in the ranking who has chosen that major. The major with highest number of top students choosing it would be the highest in the ranking (Columns (3) and (4) of

Table 2.3). The following equations show the naive model of major popularity:

$$Avg\ Rank_{jt} = \alpha_1 \log(Annual\ Salary)_{jt} + \alpha_2 Unemp_{jt} + \alpha_3 Growth_{jt} + \epsilon_{jt} \quad (2.1)$$
$$Top\ Students_{jt} = \beta_1 \log(Annual\ Salary)_{jt} + \beta_2 Unemp_{jt} + \beta_3 Growth_{jt} + \epsilon_{jt} \quad (2.2)$$

Table 2.3: Determinants of Major Popularity

	Average Rank of Students		Number of Students Ranked < 10000	
	(1)	(2)	(3)	(4)
Log Annual Salary	-5548.4	-7522.4*	47.47***	52.89***
	(3500.1)	(3718.8)	(14.04)	(15.09)
Unemployment Rate	1014.5***	1127.3***	-4.188***	-4.283***
	(260.4)	(268.4)	(1.045)	(1.089)
Salary Growth Rate		7144.6		-13.48
		(3921.9)		(15.91)
Constant	148424.1*	179834.2**	-682.2**	-776.0**
	(64274.2)	(67934.2)	(257.9)	(275.7)
Observations	571	549	571	549
R^2	0.029	0.040	0.044	0.046

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2.3 shows that labor market variables such as income and unemployment rate affects the popularity of a major. Here I assume that students know about the labor market variables and they form their expectations based on the available data. Columns (1) and (2) show that majors with higher income and lower unemployment rate are chosen by students higher in ranking. Same conclusion can be drawn by looking at the columns (3) and (4). The coefficients can be interpreted as if careers associated with major J earn one percent more annual salary than major J', J is on average chosen by about 50 more students in the top decile. Also, if the unemployment rate associated with major J is one percent lower than major J', J is on average chosen by four more students in the top decile. Both results show that majors in which currently are better in terms of labor market variables are chosen more by top students.

The OLS results show that labor market variables are important factors for a student

who is deciding what major to study in the university. Based on this result we can use revealed preferences approach to estimate a model of major choice using the individual data. I assume the probability of choosing major j by individual i to be:

$$Pr(y_i = j|x_i) = \frac{exp(x_i\beta_j)}{\sum_{k=1}^{J} exp(x_i\beta_k)} \tag{2.3}$$

By choosing $J = Humanities$ as the base category, logarithm of relative probabilities can be written as:

$$\eta_{ij} = \ln\left(\frac{Pr(y_i = j)}{Pr(y_i = J)}\right) = x_i\beta_j \tag{2.4}$$

Variables such as Age, Gender, High School GPA and City of residence are included in x_i. The trick used to identify the role of earnings in the probability of choosing a major is to use narrower categories for income compared to broader available choice categories. I assume that student chooses among ISCED97 codes which include 17 categories such as Humanities, Engineering, Law ..., while I use the major specific earnings based on SCO88 coding which includes 351 unique values for different majors. This allows me to have within group variation for earnings and to be able to add earnings to Equation 2.4. The Table 2.4 shows the estimation of Equation 2.4 using a sample of our data. The table is similar to Table 10 in Arcidiacono et al. [2010].

These result shows that girls are more likely to choose Humanities over Engineering, Architecture or Health. Students from small cities are less likely to choose Business and the probability that a high ability students chooses Engineering, Architecture and Health is higher than him choosing Humanities. It also shows that earnings affect the probability of choosing different majors. Although it is hard to interpret the coefficients, significant result shows that with change in the earnings, probability of choosing different major changes. This is another evidence for the importance of labor market outcome for major choice.

So far, we haven't used the main feature of the dataset which is the ordered preferences of majors. For this part, similar to Arcidiacono et al. [2010] and Zafar [2013] I will estimate a

Table 2.4: Multinomial Logit Estimation of Major Choice

	ISCED97			
	Business & Addministration	Engineering	Architecture & Building	Health
Female	-0.209	-2.088***	-1.458**	-1.762**
	(-0.39)	(-3.78)	(-2.73)	(-2.89)
Mid Size Cities	-0.482	-0.124	-0.944	0.819
	(-0.84)	(-0.21)	(-1.64)	(1.22)
Small Cities & Villages	-1.336*	-0.614	-1.430*	1.202
	(-2.00)	(-0.89)	(-2.15)	(1.56)
High School GPA	0.118	0.405***	0.314***	1.443***
	(1.47)	(4.71)	(3.82)	(8.85)
Age	-0.164	0.00153	-0.0360	0.671*
	(-0.69)	(0.01)	(-0.15)	(2.31)
Log SCO88 Salary	8.785***	16.41***	9.957***	17.33***
	(8.35)	(13.37)	(9.08)	(13.67)
Constant	-160.9***	-311.9***	-187.2***	-362.6***
	(-7.90)	(-13.04)	(-8.80)	(-14.33)
Observations	728			

Humanities is the base category. Other categories are omitted to improve readability.

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

rank ordered logit model of major choice using the data in hand. Let $r_i = (r_{i1}, r_{i2}, ..., r_{i100})'$, where r_{im} is the major/university (course) that student i has ranked as the mth highest pair of 100 majors/universities. If v_{ijs} denotes the indirect utility function of student i from choosing major j in school s ($r_{im} = \{j, s\}$) and assuming that students choose their rankings to maximize their utility, probability of observing r_i for student i, would be:

$$Pr(r_i) = Pr(v_{ir_{i1}} > v_{ir_{i2}} ... > v_{ir_{i100}}) \tag{2.5}$$

Assuming that unobservable preferences follow a Type I extreme value distribution, the probability that student i chooses ranking r_i is:

$$Pr(r_i) = \prod_{m=1}^{99} \frac{exp(v_{ir_{im}})}{\sum_{l=m}^{100} exp(v_{ir_{il}})} \tag{2.6}$$

And the log likelihood for the data would be:

$$L = \sum_{i=1}^{N} \log[Pr(r_i)] \tag{2.7}$$

Table 2.5 shows the maximum likelihood estimation of the rank ordered model by using 2015 data. Dependent variable of the model would be the row number of the course in the student's list. To be more clear, higher ranked courses will have a smaller row number assigned to them. Main independent variables are SCO88 average salary and unemployment rate extracted from the 2015 LFS and 2015 HIES. I have also constructed a variable for studying in own state or another state. As column (1) of Table 2.5 shows, higher earnings and higher employment rate are associated with lower row number of the course in student's list. Also, students list the courses in their own state in rows with smaller numbers which means they prefer to stay in close to their hometown. Column (2) shows the fact that studying close to family is more important for girls compared to boys. The positive coefficient on *Female* variable states the fact that girls list the courses out of their state in rows with larger numbers. Column (3) shows that students from lower socioeconomic status are more willing to stay in their own state to study and larger coefficient on *Small Cities and Villages* shows its importance for poorest students. Note that better universities are mostly in large cities so column (3) might be an evidence of credit constrained students who cannot afford living far from their family.

	(1) Row	(2) Row	(3) Row
Log Annual Salary SCO88	-0.194***	-0.194***	-0.195***
	(0.00147)	(0.00147)	(0.00150)
Unemployment	0.0288***	0.0288***	0.0287***
	(0.000121)	(0.000121)	(0.000124)
Own State	-0.490***	-0.508***	-0.674***
	(0.00185)	(0.00274)	(0.00304)
Female		0.0389***	
		(0.00250)	
OwnState × Female		-0.435	
		(.)	
Mid Size Cities			0.340***
			(0.00298)
Small Cities & Villages			0.394***
			(0.00384)
Own State × Mid Size Cities			0
			(.)
Own State × Small Cities & Villages			-0.0406
			(.)
Observations	3713118	3713118	3585197

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2.5: Rank Ordered Logit Estimation of Major Choice

Large number of observations cause the standard errors to be very low and the coefficient to be estimated with high precision. This is an important advantage of this study over other studies in the literature which are mostly based on surveys on a limited number of students. Another advantage is the school dimension of dataset which is rarely mentioned in the literature. As results show, students care about the labor market outcome and they also prefer to study close to their hometown. Most of the studies in the literature are done in the scale of one college or one university and they are unable to capture the importance of location of the school.

To make sure that the results are not sensitive to the major categories, I will reestimate the model using ISCED97 broad categories. The results are the same and are shown in Table 2.6.

	(1) Row	(2) Row	(3) Row
Log Annual Salary ISCED97	-0.147***	-0.147***	-0.145***
	(0.00264)	(0.00264)	(0.00266)
Unemployment	0.0343***	0.0343***	0.0342***
	(0.000121)	(0.000121)	(0.000123)
Own State	-0.469***	-0.492***	-0.644***
	(0.00180)	(0.00269)	(0.00296)
Female		0.450***	
		(0.00242)	
Own State × Female		-0.000221	
		(.)	
Mid Size Cities			0.323
			(.)
Small Cities & Villages			0.383***
			(0.00374)
OwnState × Mid Size Cities			-3.47e-09
			(0.00291)
OwnState × Small Cities & Villages			-0.0375
			(.)
Observations	3862037	3862037	3731132

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2.6: Rank Ordered Logit Estimation of Major Choice Using ISCED97 Categories

2.4 Conclusion

Using the unique dataset on major/university choices of Iranian students who took the nationwide university entrance exam, I estimated a rank ordered logit model of major choice. I showed that labor market variables, specifically earnings and unemployment play a signifi-

cant role in choice of majors by students. Labor market variables were extracted from Labor Force Survey (LFS) and Household Income and Expenditure Survey (HIES) and were linked to different majors by International Standard Classification of Education 1997 (ISCED97) and Selected Characteristics of Occupations 1988 (SCO88). The model showed that students prefer majors with higher expected income and expected employment rate in the sense that they list these majors higher in their ordered ranking of majors.

Another finding of this paper is related to theories suggesting that many people might only care about the school and not the major. It can have several explanations, for example prestige of some schools might be one reason. Credit constraints of family or the cultural barriers might also play role for those students who prefer to stay in their hometown even at the price of studying a major that they are not very interested in. I provided evidence for these theories by including the dummy variable for studying in one's own state in my model. Significance of this variable and its interactions with gender and SES variables shows that there are such cultural barriers and credit constraints.

We are conducting a survey on students who took the Concurs in 2010 asking them about earnings and labor market status. This will give us the opportunity to run tests on how well people did on foreseeing future outcome and also let us run a regression discontinuity model to calculate returns to different majors.

Appendix

2.5 Descriptive Results

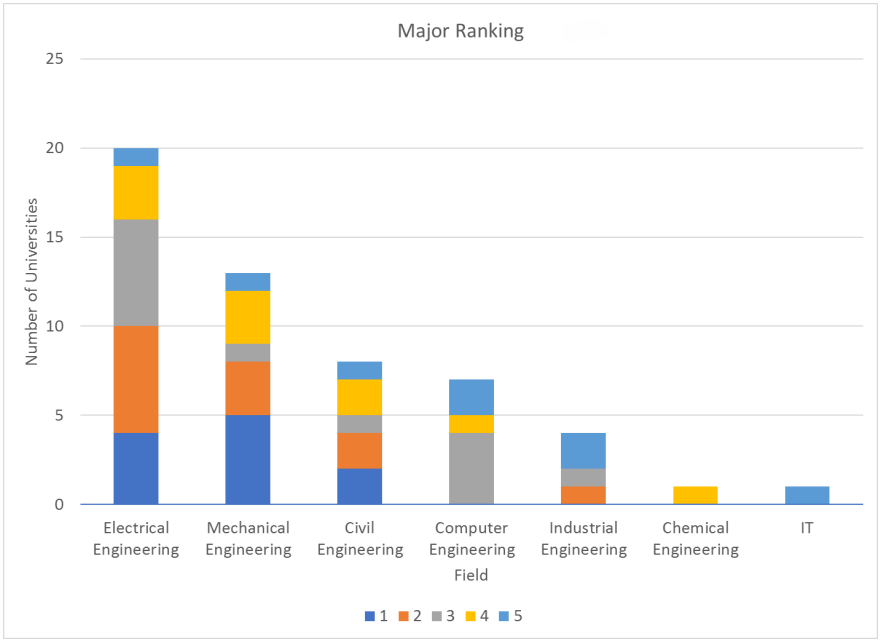
For finding the most popular majors, for each university, I have sorted the ranking of last person who has filled the seats of that major in that specific university. This has been done for all 12 universities for 2013 and 2015 Concours and the results are shown in Figure 2.2.

As shown in Figure 2.2a, Electrical Engineering was by far the most popular major in 2013. Different fields of electrical engineering (Electronics, Communication, Power) were the first major to be filled in 4 universities, filled second in 6 and filled third in 6 universities. Mechanical Engineering, Civil Engineering and Computer Engineering were the next popular majors in 2013.

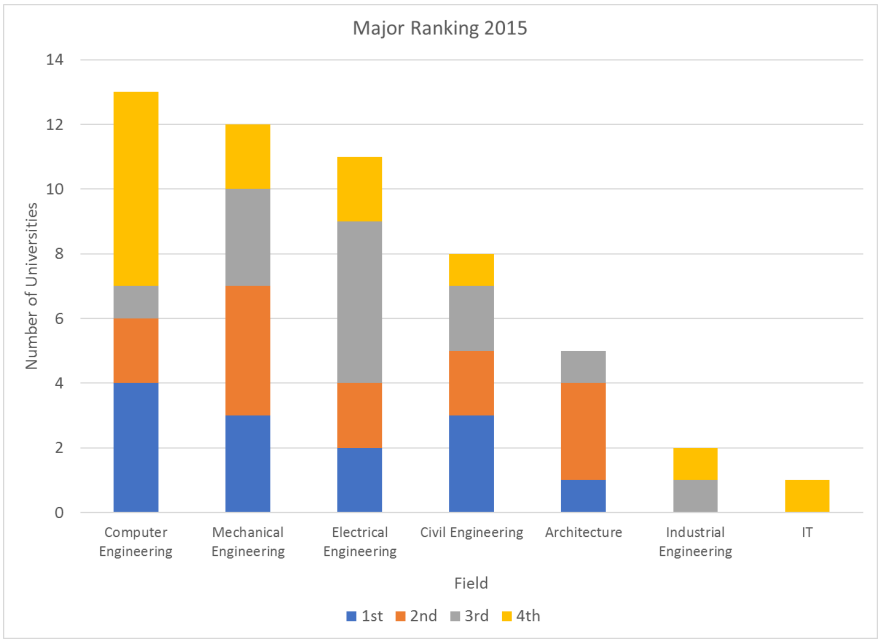
Comparing Figure 2.2a and Figure 2.2b shows that 4 most popular majors are the same as 2013 in 2015 but the rankings has changed. Electrical Engineering which was the most popular major in 2013 by far, became third in ranking in 2015. On the other hand Computer Engineering which was the fourth popular major in 2013, became the most popular one two years later. It seems that students have responded to the growing labor market demand for computer engineers.

The interesting fact is that similar pattern can be seen among the UCLA applicants too. Figure 2.3 shows the share of transfer applicants to different UCLA engineering majors for the past 15 years. As can be seen, Computer Science has became very popular in recent years, while Electrical Engineering and Civil Engineering are not very popular anymore.

Intuitively, it seems that students respond to the signals from the labor market and those majors which related to growing occupations become more popular among the students. It is important to see whether this major ranking is the same for average students too and whether average students respond to labor market trends similar to top students or not. This raises the question that to what extent people care about the labor market outcome when they are choosing their major? And also how responsive major choice is to labor market changes? These questions cannot be answered with this limited public dataset and



(a) 2013



(b) 2015

Figure 2.2: Majors Ranking

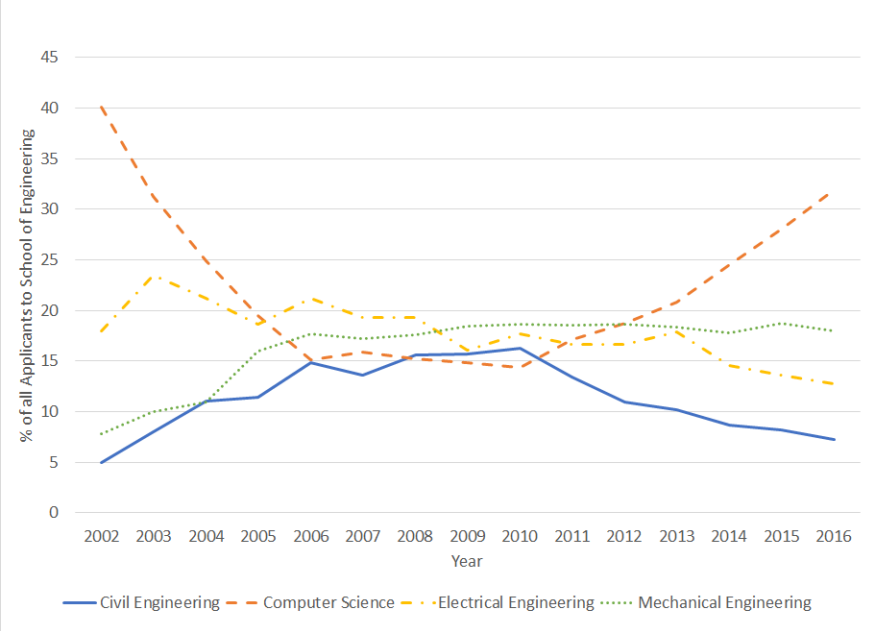


Figure 2.3: Evolution of Share of All Applicants to Engineering School
Source: UCLA Undergraduate Admission Profile of Transfer Applicants

needs access to the administrative one.

If salaries are added to Figure 2.2a we can see that relationship between the most popular major and most paid occupations is not very clear. Using the Consumer Expenditure and Income Survey and averaging the income of workers who graduated with a degree in one of 18 majors, salaries associated to each majors' graduates has been found. Figure 2.4 shows ranking of majors with the associated salary. It can be seen that although some majors are very popular among students but the graduates are not among the top paid workers. For example, Mechanical Engineering and Civil Engineering are second and third in the majors ranking, but the graduates are paid less than Computer or Chemical Engineering graduates, which are not among the top majors.

Next figure provides evidence for what is referred to as "school prestige" in the major choice literature. For each major, universities have been sorted based on the ranking of last person who got a seat in that university. As can be seen in Figure 2.5, Sharif University is the most prestigious school in Iran. All majors that are offered in Sharif are filled first among all the universities. Another important observation is that top five universities are

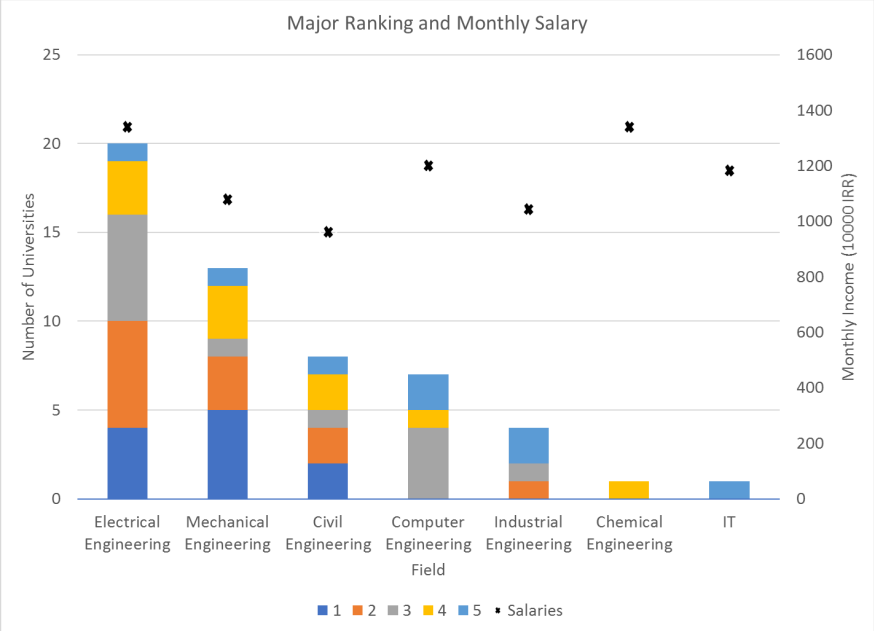


Figure 2.4: Majors Ranking and Monthly Salary 2013

all in Tehran which suggests that quality of schools might be very unequally distributed in the country.

Figure 2.6 shows the share of non-resident students in each school in the sample. This is an important evidence suggesting that people might choose schools rather than majors. Share of non-resident students in universities in Tehran shows that students from other cities are willing to leave their home town for the capital only if they got accepted to Sharif University. However, this is not true about the universities in cities other than Tehran. The share of non-residents in those universities are relatively high with most of non-residents coming from Tehran. This suggests that families from cities other than Tehran might face some kind of a credit constraints that the students would prefer to stay at their hometown.

Figure 2.5 and Figure 2.6 put emphasis on an important question in the literature: Do people choose major or they choose the school? This question can be answered very clearly using the data on 100 ordered choices of students. To be more clear on this, if one student has listed one major in different universities as her top choices, it shows that she gives priority to the major. On the other hand, if one student has listed all majors in one university at the top of his list, he is definitely choosing school and not the major. The difference here

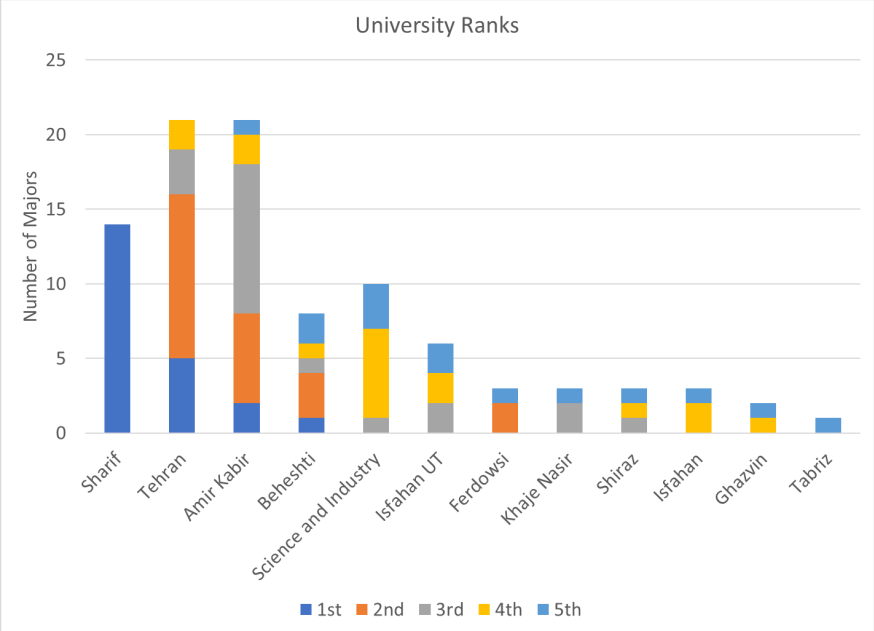


Figure 2.5: School Ranking

might be very important because in the literature it is always assumed that the major that student is studying in college is his first choice and other majors are his alternatives.

This hypothesis should also be tested that, does the share of students who choose major and the students who choose the school differs among top students and average students. It seems possible that top students who can choose among more options, choose their desired major but average students choose the school and not the major.

In this section, public data on Mathematics and Physics Concours were used to provide some descriptive results which can be suggestive for what interesting questions to answer using the administrative dataset.

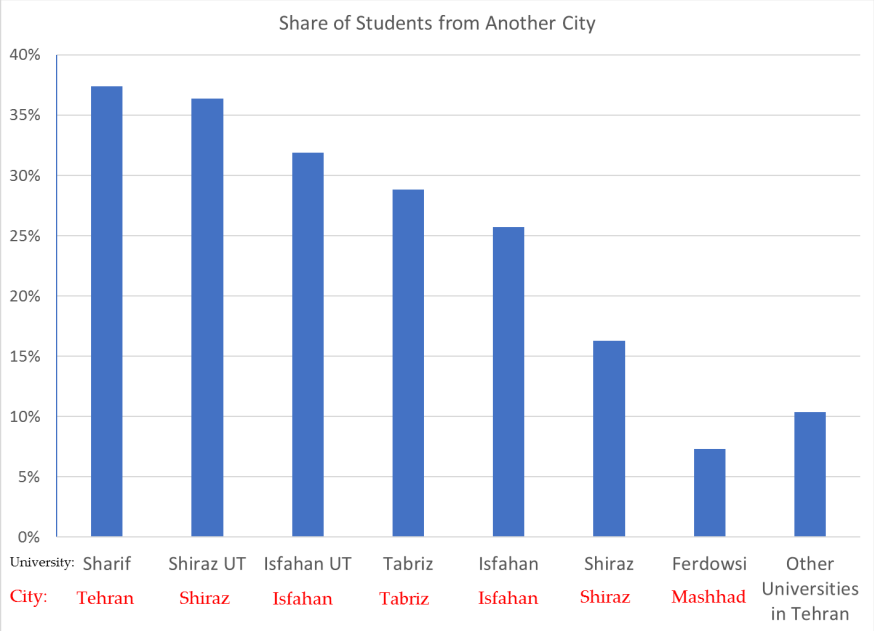


Figure 2.6: Share of Non-Resident Students in Different Schools

2.6 Standard Classifications

Category (ISCED97)	Major Example (SCO88)
Teacher training and education science	Teacher Training
Arts	Cinema
Humanities	Foreign Languages
Social and Behavioural Science	Economics
Journalism and Information	Media
Business and Administration	Accounting
Law	Law
Life Sciences	Biology
Physical Sciences	Geomorphology
Mathematics and Statistics	Statistics
Computing	Computer Science
Engineering and Engineering Trades	Electrical Engineering
Manufacturing and Processing	Mining Engineering
Architecture and Building	Civil Engineering
Agriculture, Forestry and Fishery	Agriculture
Veterinary	Veterinary
Health	Medicine

Table 2.7: List of Categories and Major Examples

CHAPTER 3

Using Neural Networks to Predict Length of Study at UCLA

3.1 Introduction

The admissions process is one of the most important yet time consuming processes that is associated with university management. Given the limited amount of capacity to intake students due to resource constraints, such as a limited number of classrooms, professors, and so forth, officials need to select whom to admit carefully. The admissions officers need to minimize resource waste by admitting only the students whom they think have the academic and social background to withstand the rigors of higher education. They also need to prioritize selection of students whom they expect will be able to graduate early and thereby open space for new cohorts.

Using a dataset on UCLA students who were admitted in fall 1983 and studied for at least two quarters, we attempt to identify the characteristics of students who are unlikely to graduate within six years of admission. We chose six years because it is the longest acceptable duration for a student to study at UCLA. In this study, we apply Machine Learning methods, specifically Neural Networks, to predict two outcomes: i) the number of quarters that it will take the student to graduate and ii) whether or not the student will be able to graduate within six years. The outcome of the models are normalized or compared to linear regression model outcomes to determine the value added of using Machine Learning techniques.

3.2 Data

We are using a subset of a administrative dataset on UCLA students who were first admitted in the Fall of 1983. Because of some data issues, we restrict our data to those students who graduated between 2 and 100 quarters which makes it between half a year to 25 years.

The restricted data includes more than 4700 students. The average number of quarters for graduation is almost 16 quarters or 4 years and the median is 14 quarters or 3 years and a half. Share of the transfer students is almost 45 percent which gives some sense to the average graduation time.

3.3 Methodology

3.3.1 Number of Quarters Prediction

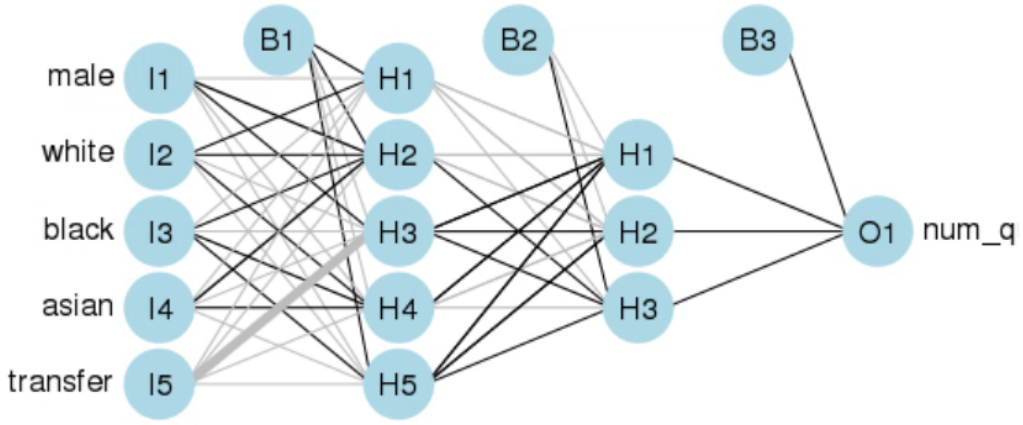
We are using a Neural Network approach to predict how many quarters it takes a UCLA student to graduate depending on his or her characteristics. Further we are going to explore what are the characteristics of a student who was not able to graduate in 6 years after being admitted to UCLA.

We use the data described in the previous section to run the model. The first model is supposed to predict number of quarters that it takes a student with certain characteristics to graduate. The Neural Network is trained by 60 percent of the sample, then the coefficients of the model are determined by 20 percent of the sample in the cross validation stage. Finally, the model is tested with the remaining 20 percent of the sample.

The first model that we run is designed to predict number of quarters for graduation by the following binary variables: Male, White, Black, Asian and Transfered. The Neural Network has two hidden layers, first one with 5 and second one with 3 active nodes. The structure of this network is shown in Figure 3.1. The mean squared error from running this model on the test subsample is 77.11 which is higher than the same measure obtained from a linear regression model (58.99). Comparing these two numbers and getting a higher number

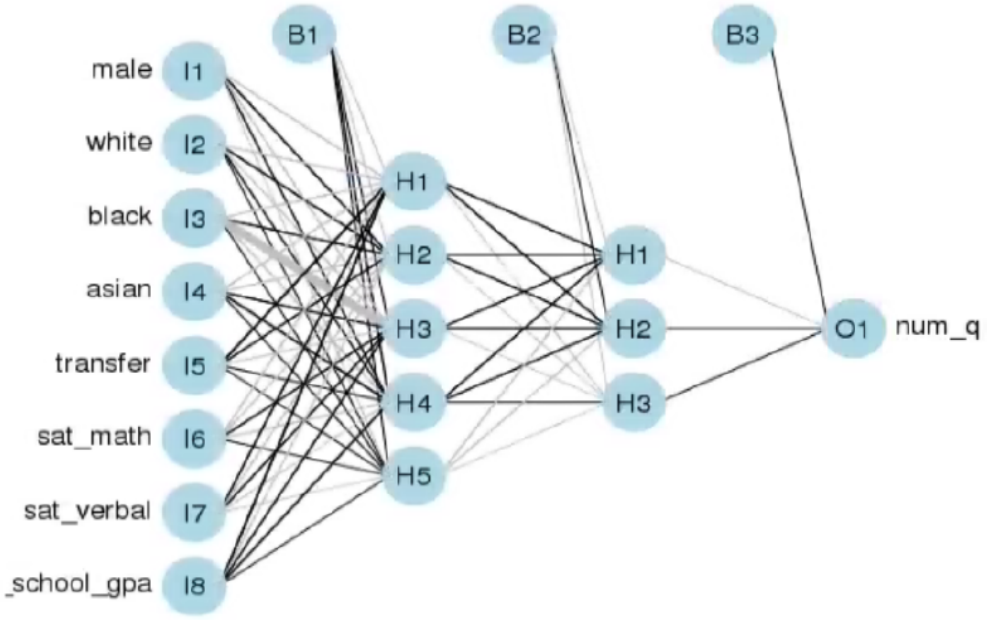
from the neural network shows that we are not gaining from using the machine learning techniques to predict number of quarters. The problem might be because of the restricted sample that we are using or the parameters of the model need to be changed. Changing the parameters does not help decreasing the mean squared errors of the model. Using the whole data or restricting the subset in hand might improve the performance of the model.

Figure 3.1: Neural Network for Predicting Number of Quarters for Graduation with Two Hidden Layers (5,3).



In the next step, we will add more variables such as SAT Verbal, SAT Math and high school GPA. These variables are not binary variables and we expect them to improve the prediction power of the model. The structure of this model is shown in Figure 3.2. About 1800 students don't have SAT scores so we have to remove them before running the model. The mean squared error of this model is 71.18 which again is higher than the mean squared error from linear regression (55.16). In the linear regression SAT score variables are found statistically insignificant and the high school GPA is only significant in 10 percent level of confidence.

Figure 3.2: Neural Network for Predicting Number of Quarters for Graduation with Two Hidden Layers (5,3).



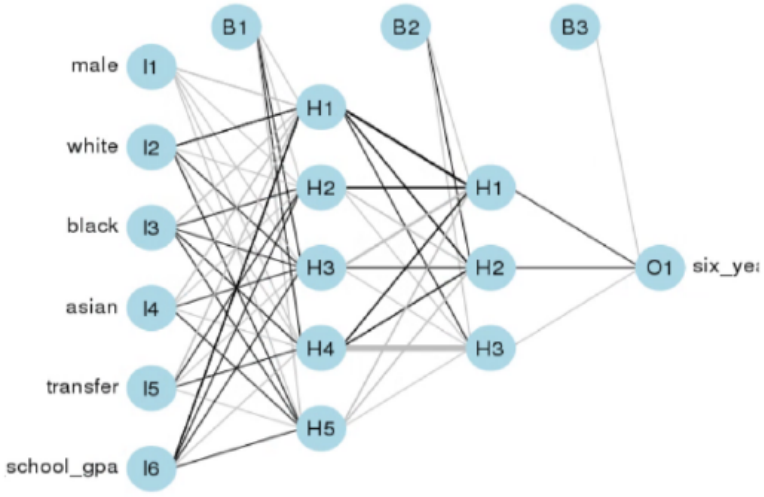
By looking at the results from the two models in this subsection, it seems that Machine Learning is not able to do a good job in predicting the number of quarters for graduation. Linear regression gives a better fit than machine learning and this might be because of the data limitation. In the subset of data that is used here it seems that Machine Learning cannot give a good prediction of how many quarters it takes a student who was admitted in Fall of 1983 to graduate.

3.3.2 Survival Model

After running the previous model, we move to form a survival model. In this model we assume that those who finish their study before the sixth year are the survivors and try to predict the probability of surviving using data on characteristics of the students. Using the results from the previous section, we only include the demographics variables plus transfer and the high school GPA for running the survival model. The structure of the model is

shown in Figure 3.3.

Figure 3.3: Neural Network Structure of the Survival Model



Mean squared error of this model is 0.081022 which is higher than the mean squared error from regressing the survival variable on the predictors (0.075295). But here MSE is not a useful measure and we care about the share of the survived people that the model predicts correctly or the share of the people who fail to graduate before six years (deaths). In our sample, number of people who graduate before sixth year is almost 10 times as those who graduate after sixth year. So using the regular measures of accuracy might be misleading. In these cases which the dataset is not balanced in terms of numbers deaths and survivors, even predicting survival for everyone will give a high percentage for correct predictions.

Looking at the accuracy curve Figure 3.4 shows the problem that was mentioned in the last paragraph. As it can be seen from the graph, even putting the threshold cutoff equal to 0 will generate a high accuracy in terms of share of correct predictions and this is only because of the fact that the number of survivals are much higher than the number of deaths.

Figure 3.4: Accuracy Curve for the Survival Neural Network

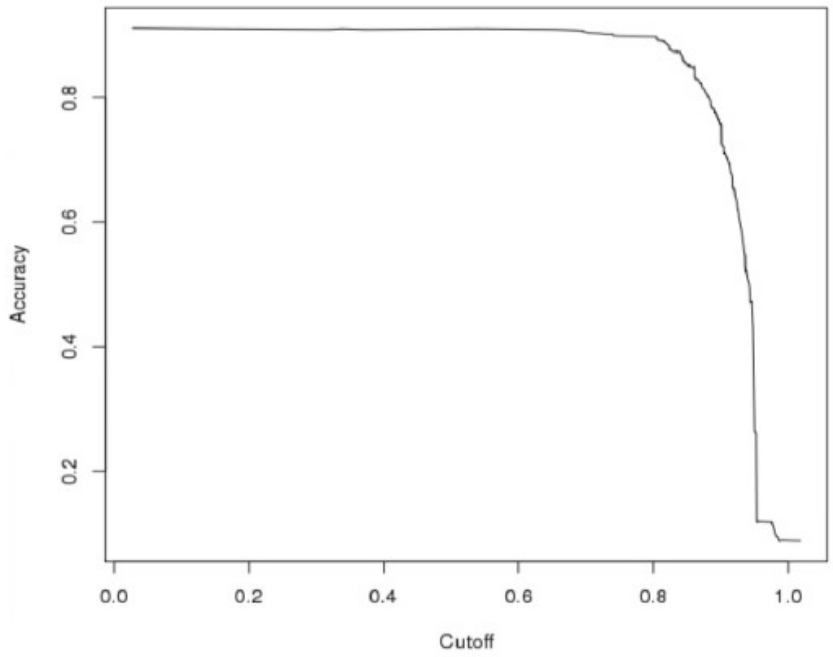
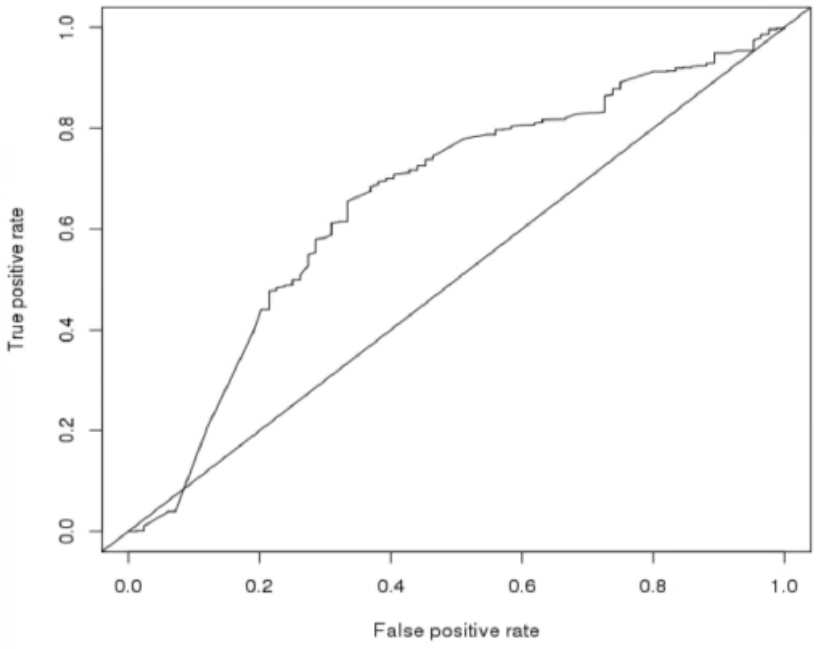


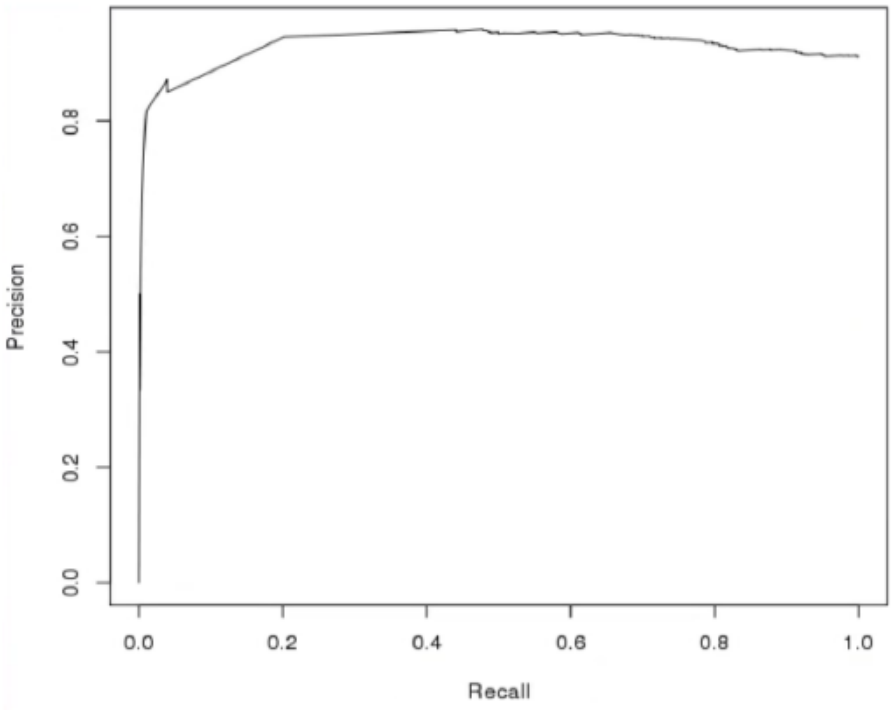
Figure 3.4 shows that changing the cutoff does not help the prediction of model and the True Positives are the dominant no matter what the cutoff is. Better measures of how good the model is predicting who graduates before sixth year are Relative Operating Characteristic (ROC) curve and Recall-Precision (RP) curve. In the ROC curve true positive rate is graphed against the false positive rate. The area under curve in ROC curve is equal to probability that the model ranks a random positive instance higher than a random negative one. The ROC curve for our survival model is shown in Figure 3.5. The area under curve is calculated to be 0.659.

Figure 3.5: Relative Operating Characteristic Curve



As can be seen from the graph, predicting survival is not random but it is not also being done perfectly by model. The area under the curve is less than an amount that the model could be trusted. The problem can be coming from the data issues that were mentioned before or insufficient number of variables are included. The RP curve is also shown in Figure 3.6.

Figure 3.6: Recall Precision Curve



3.4 Conclusion

In this study we used machine learning techniques, specifically a Neural Network approach, to predict the number of quarters that it takes a student with certain characteristics to graduate from UCLA. We used a subset of students who got admitted to UCLA in Fall 1983 either as a transfer student or a non-transfer student. The subset includes over 4700 students with different demographic and academic characteristics. We first used the Neural Network to predict number of quarters needed for graduation. By comparing the Mean Squared Errors of the network and the one from linear regression, we concluded that using machine learning is not increasing our prediction accuracy. Next we defined a survival model for our study, in such those who did graduate before sixth year were survivors and those who couldn't were the failures in our model. This model's performance was not also great mainly because of the fact that the dataset is unbalanced in the sense that number of failures are much less than number of successes. The area under curve of the relative operating characteristic curve was

calculated to be 0.66 which is not a very high number. Limited accuracy of the models here might be all because of the restrictions of data and the fact that only a subset of original dataset is used. Using the whole dataset might give us a very different result in terms of prediction and accuracy but that requires a lot of data cleaning.

Bibliography

Atila Abdulkadiroglu, Parag Pathak, Alvin E. Roth, and Tayfun Sonmez. Changing the Boston school choice mechanism. Technical report, National Bureau of Economic Research, 2006.

Atila Abdulkadiroğlu and Tayfun Sönmez. School choice: A mechanism design approach. *American economic review*, 93(3):729–747, 2003.

Atila Abdulkadiroğlu, Nikhil Agarwal, and Parag A. Pathak. The welfare effects of coordinated assignment: Evidence from the New York City high school match. *American Economic Review*, 107(12):3635–89, 2017.

Nikhil Agarwal and Paulo Somaini. Demand Analysis using Strategic Reports: An application to a school choice mechanism. *Econometrica*, 86(2):391–444, 2018.

Kehinde Ajayi and Modibo Sidibe. An empirical analysis of school choice under uncertainty. Technical report, Boston University Working paper, http://people. bu. edu/kajayi/research. html, 2015.

Joseph G. Altonji and Seth D. Zimmerman. The Costs of and Net Returns to College Major. Technical Report 23029, January 2017. URL http://www.nber.org/papers/w23029. DOI: 10.3386/w23029.

Donald WK Andrews and Xiaoxia Shi. Inference based on conditional moment inequalities. *Econometrica*, 81(2):609–666, 2013.

Peter Arcidiacono. Ability sorting and the returns to college major. *Journal of Econometrics*, 121(1-2):343–375, 2004. ISSN 0304-4076. URL http://econpapers.repec.org/article/eeeeconom/v_3a121_3ay_3a2004_3ai_3a1-2_3ap_3a343-375.htm.

Peter Arcidiacono, V. Joseph Hotz, and Songman Kang. Modeling College Major Choices using Elicited Measures of Expectations and Counterfactuals. Technical Report 15729, February 2010. URL http://www.nber.org/papers/w15729. DOI: 10.3386/w15729.

Georgy Artemov, Yeon-Koo Che, and Yinghua He. Strategic 'Mistakes': Implications for Market Design Research. Technical report, mimeo, 2017.

Rachel Baker, Eric Bettinger, Brian Jacob, and Ioana Marinescu. The Effect of Labor Market Information on Community College Students' Major Choice. Technical Report 23333, April 2017. URL http://www.nber.org/papers/w23333. DOI: 10.3386/w23333.

Magali Beffy, Denis Fougère, and Arnaud Maurel. Choosing the Field of Study in Post-secondary Education: Do Expected Earnings Matter? *The Review of Economics and Statistics*, 94(1):334–347, May 2011. ISSN 0034-6535. doi: 10.1162/REST_a_00212. URL http://dx.doi.org/10.1162/REST_a_00212.

Eric Budish and Estelle Cantillon. The multi-unit assignment problem: Theory and evidence from course allocation at Harvard. *American Economic Review*, 102(5):2237–71, 2012.

Hector Chade and Lones Smith. Simultaneous search. *Econometrica*, 74(5):1293–1307, 2006.

Li Chen. University admission practices–Ireland. *MiP Country Profile*, 8, 2012.

Monique De Haan, Pieter A. Gautier, Hessel Oosterbeek, and Bas Van der Klaauw. The performance of school assignment mechanisms in practice. 2015.

Torben Drewes and Christopher Michael. How do students choose a university?: an analysis of applications to universities in Ontario, Canada. *Research in Higher Education*, 47(7): 781–800, 2006.

Lester E. Dubins and David A. Freedman. Machiavelli and the Gale-Shapley algorithm. *The American Mathematical Monthly*, 88(7):485–494, 1981.

Gabrielle Fack, Julien Grenet, and Yinghua He. Beyond Truth-Telling: Preference Estimation with Centralized School Choice and College Admissions. *American Economic Review*, 109(4):1486–1529, 2019.

David Gale and Lloyd S. Shapley. College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15, 1962.

Guillaume Haeringer and Flip Klijn. Constrained school choice. *Journal of Economic theory*, 144(5):1921–1947, 2009.

Justine Hastings, Thomas J. Kane, and Douglas O. Staiger. Heterogeneous preferences and the efficacy of public school choice. *NBER Working Paper*, 2145:1–46, 2009.

Justine Hastings, Christopher A. Neilson, and Seth D. Zimmerman. The effects of earnings disclosure on college enrollment decisions. Technical report, National Bureau of Economic Research, 2015a.

Justine S. Hastings, Christopher A. Neilson, and Seth D. Zimmerman. Are Some Degrees Worth More than Others? Evidence from college admission cutoffs in Chile. Technical Report 19241, July 2013. URL http://www.nber.org/papers/w19241. DOI: 10.3386/w19241.

Justine S. Hastings, Christopher A. Neilson, Anely Ramirez, and Seth D. Zimmerman. (Un)Informed College and Major Choice: Evidence from Linked Survey and Administrative Data. Technical Report 21330, July 2015b. URL http://www.nber.org/papers/w21330. DOI: 10.3386/w21330.

Martin Hällsten. The structure of educational decision making and consequences for inequality: A Swedish test case. *American Journal of Sociology*, 116(3):806–54, 2010.

Adam Kapor, Christopher A. Neilson, and Seth D. Zimmerman. Heterogeneous beliefs and school choice mechanisms. Technical report, National Bureau of Economic Research, 2018.

Lars J. Kirkeboen, Edwin Leuven, and Magne Mogstad. Field of study, earnings, and self-selection. *The Quarterly Journal of Economics*, 131(3):1057–1111, 2016.

Lars Johannessen Kirkebøen. Preferences for lifetime earnings, earnings risk and nonpecuniary attributes in choice of higher education. Technical report, Discussion Papers, 2012.

Shengwu Li. Obviously strategy-proof mechanisms. *American Economic Review*, 107(11): 3257–87, 2017.

Margaux Luflade. The value of information in centralized school choice systems. *Duke University*, 2018.

Claude Montmarquette, Kathy Cannings, and Sophie Mahseredjian. How do young people choose college majors? *Economics of Education Review*, 21(6):543–556, December 2002. ISSN 0272-7757. doi: 10.1016/S0272-7757(01)00054-1. URL http://www.sciencedirect.com/science/article/pii/S0272775701000541.

Alvin E. Roth. The college admissions problem is not equivalent to the marriage problem. *Journal of economic Theory*, 36(2):277–288, 1985.

Perihan Ozge Saygin. Gender differences in preferences for taking risk in college applications. *Economics of Education Review*, 52:120–133, 2016.

Ralph Stinebrickner and Todd Stinebrickner. Academic Performance and College Dropout: Using Longitudinal Expectations Data to Estimate a Learning Model. Technical report, 2013. URL https://ideas.repec.org/p/uwo/hcuwoc/20135.html.

Matthew Wiswall and Basit Zafar. How Do College Students Respond to Public Information about Earnings? *Journal of Human Capital*, 9(2):117–169, June 2015. ISSN 1932-8575. doi: 10.1086/681542. URL http://www.journals.uchicago.edu/doi/abs/10.1086/681542.

Basit Zafar. College Major Choice and the Gender Gap. *Journal of Human Resources*, 48 (3):545–595, July 2013. ISSN 0022-166X, 1548-8004. doi: 10.3368/jhr.48.3.545. URL http://jhr.uwpress.org/content/48/3/545.